Por uma Análise de Conteúdo Mais Fiável

Jorge Ávila de Lima[1]

Resumo

A análise de conteúdo é uma técnica muito invocada e utilizada nas ciências sociais e humanas, incluindo as ciências da educação, mas raramente desenvolvida de forma adequada. No presente texto, defende-se a necessidade de uma abordagem mais sistemática e rigorosa a esta análise e propõem-se soluções técnicas para algumas das principais situações com que os analistas de conteúdo se deparam. O texto inclui recomendações metodológicas relativas a diversos aspetos deste tipo de trabalho e uma lista de verificação que permite aos investigadores determinarem em que medida as suas análises se pautam por critérios de rigor neste domínio.

Palavras-chave: análise de conteúdo; fiabilidade; *kappa* de Cohen

Introdução

A análise de conteúdo é uma técnica que permite a classificação de material, reduzindo-o a uma dimensão mais manejável e interpretável, e a realização de inferências válidas a partir desses elementos (Weber, 1990). Pela sua abrangência, a definição de Kolbe e Burnett (1991, p. 243) é particularmente adequada: "content analysis is an observational research method that is used to systematically evaluate the symbolic content of all forms of recorded communications". Quando aplicada ao material escrito, o objetivo básico desta análise consiste em reduzir as muitas palavras de um texto a um pequeno conjunto de categorias de conteúdo (Bardin, 1995). À semelhança do que fazem os estatísticos com a análise de dados quantitativos, também os analistas de conteúdo procuram sintetizar e reduzir a quantidade de informação disponível, para chegarem a uma interpretação das principais tendências e padrões presentes nos seus dados.

1 Professor Associado com Agregação
Departamento de Ciências da Educação, Universidade dos Açores. E-mail: javilalima@hotmail.com

10.14195/1647-8614_47-1_1

Existem muitas modalidades de análise de conteúdo propostas na literatura publicada e os objetivos dos pesquisadores neste domínio variam bastante. Enquanto alguns a usam meramente com o intuito de classificar a informação recolhida de acordo com uma estrutura que sintetize as tendências gerais presentes nos dados, outros propõem-se captar a "verdadeira" estrutura de significado escondida por detrás desses dados.

Infelizmente, no mundo académico, este tipo de análise tem sido pensado, muitas vezes, sobretudo pela negativa: parece tratar-se de tudo o que se faz com os dados que não consista em análise quantitativa. Qualquer comentário feito a um relato realizado por um informador, qualquer seleção e apresentação de um excerto de um documento ou de uma transcrição de uma entrevista parece merecer o título de "análise de conteúdo". Do ponto de vista metodológico, isto é muito insuficiente.

Por vezes, os esforços para superar esta situação e dotar a análise de conteúdo de uma maior cientificidade são encarados com ceticismo e mesmo com uma oposição determinada (ver, por exemplo, Rocha & Deusdará, 2005). No entanto, as críticas realizadas a este propósito parecem basear-se mais na descrença da possibilidade de se construir verdadeiro conhecimento científico, seja qual for o tipo de dados, do que propriamente em qualquer objeção específica relativamente aos métodos propostos para este tipo de análise.

A questão da análise de conteúdo de dados de investigação tem sido tratada em alguns textos publicados sobre a matéria, em língua portuguesa (Bardin, 1995; Esteves, 2006; Vala, 1986). Existem, contudo, aspetos metodológicos que continuam pouco claros ou que não são sequer abordados, especialmente na grande maioria dos trabalhos de investigação que vão sendo realizados no nosso país.

O problema não é, contudo, exclusivamente nacional. Numa revisão de 200 estudos da área da comunicação de massas que utilizaram a análise de conteúdo, publicados entre 1994 e 1998, Lombard, Snyder-Duch e Bracken (2002) verificaram que apenas 69% continham alguma informação sobre a fiabilidade do trabalho de codificação realizado e que, mesmo nestes artigos, os autores apresentavam poucos detalhes sobre o modo como a análise tinha sido conduzida.

Para que seja credível, a análise de conteúdo tem de estar à altura dos melhores padrões metodológicos estabelecidos globalmente no campo da investigação científica. Por esta razão, aqueles que a realizam não podem continuar a ignorar questões tão importantes como a objetividade, a sistematização, a quantificação, a amostragem e a fiabilidade (Kassarjian, 1977; Kolbe & Burnett, 1991). Para que as inferências realizadas pelos investigadores a partir dos seus dados, com base na análise de conteúdo, sejam válidas, também importa que os procedimentos de classificação sejam consensualizados, para que diferentes pessoas possam realizar

essa classificação de forma semelhante. É igualmente essencial que o processo de análise de conteúdo seja transparente, público e verificável (Constas, 1992).

Infelizmente, existem poucas orientações disponíveis relativamente à forma como se deve atender a estas questões, o que dá origem, naturalmente, a muita inconsistência na forma como tais assuntos são abordados.

No presente artigo, procura-se fornecer recomendações metodológicas que poderão ajudar a superar algumas destas lacunas. No texto, aborda-se a técnica da análise de conteúdo, aplicada a dados obtidos em duas situações principais de pesquisa: as respostas dos inquiridos a questões abertas de questionários e as transcrições de entrevistas. Não se deve esquecer, todavia, que esta técnica também é aplicável a outros tipos de informação, decorrentes, por exemplo, da análise documental (cartas, artigos de jornal, atas, documentos históricos, discursos políticos, textos legais, diários, etc.) ou até da observação (por exemplo, notas de campo registadas por um ou mais observadores independentes). As considerações que serão feitas mais adiante reportam-se a uma análise de conteúdo de natureza *temática*, isto é, que procura identificar temas e subtemas (categorias e subcategorias) na informação a analisar. Neste artigo, distinguem-se dois cenários de aplicação da análise de conteúdo: a análise pré-estruturada e a análise aberta.

Alguma terminologia básica

No sentido de precisar a linguagem que será utilizada ao longo deste texto, recorda-se aqui alguma terminologia própria da área, devidamente clarificada por Bardin (1995), Esteves (2006) e Vala (1986), entre outros, e que tem vindo a constituir-se enquanto vocabulário específico deste tipo de análise:

- *corpus* – o conjunto dos documentos que serão objeto da análise de conteúdo;
- *"recorte"* – o processo de seleção dos segmentos de texto que serão analisados;
- *unidade de registo* – o segmento de texto que é objeto de "recorte", isto é, de seleção para análise. Geralmente, o critério de definição das unidades de registo deverá ser semântico (uma unidade com significado específico e autónomo) e não formal (por exemplo, uma palavra, uma linha, uma frase ou um parágrafo);
- *unidade de contexto* – o segmento de texto mais lato de onde é retirada a unidade de registo;
- *sistema de categorias* – conjunto de temas (categorias) que constituem conjuntos semanticamente coerentes de unidades de registo e que terão, cada um deles, um

código específico a aplicar no decurso do processo de codificação. Poderão existir (e normalmente existem) subtemas (subcategorias), que também deverão ter os respetivos códigos;

- *codificação* – processo de atribuição de códigos específicos a unidades de registo com um determinado teor semântico previamente especificado pelo investigador. Esta deve ser sempre feita ao nível mais fino (isto é, sempre que exista uma subcategoria, é o respetivo código que deve ser aplicado e não o código global da categoria-mãe);
- *codificador* ou *juiz* – pessoa que aplica os códigos presentes no sistema de categorias aos dados em análise.

Sistema de categorias

Para efetuar uma análise de conteúdo, o investigador necessitará de construir um *sistema de categorias* (também designado por vezes de "livro de códigos") para analisar a informação (ver exemplo no Anexo A[2]). Este sistema deverá conter os seguintes elementos:

- *Códigos numéricos* a utilizar para a identificação de cada categoria e subcategoria, na informação a analisar. Aconselha-se a utilização de um sistema de numeração decimal, que permita que os diferentes dígitos representem diferentes níveis (categorias, subcategorias e suas eventuais subdivisões).
- *Designação curta* ("rótulo") de cada categoria e das respetivas subcategorias;
- *Definição* por extenso de cada uma das categorias e subcategorias;
- *Exemplos típicos* de unidades de registo (normalmente, retirados dos próprios dados a analisar) que ilustram a natureza dos segmentos de informação que integram cada categoria e/ou subcategoria.

Conforme sublinhou Bardin (1995), entre outras qualidades, um bom sistema de categorias deve possuir *objetividade* e *fiabilidade*, ou seja, o mesmo material, quando analisado com base no mesmo sistema de categorias, deve ser codificado da mesma forma, mesmo quando sujeito a várias análises. Uma fiabilidade adequada garante aquilo que poderíamos designar de *objetividade intersubjetiva* dos resultados de uma

2 Trata-se de uma versão simplificada, para fins ilustrativos, de um sistema de categorias que utilizámos num projeto de investigação internacional sobre o desenvolvimento profissional dos professores.

análise de conteúdo. Dada a sua importância, este aspeto merecerá um tratamento específico e detalhado no presente texto.

Fiabilidade

Pretende-se que a investigação, para além de produzir resultados válidos, utilize instrumentos e procedimentos fiáveis. Existem três tipos de fiabilidade relevantes para a análise de conteúdo (Krippendorff, 1980, pp. 130-132): a estabilidade, a reprodutividade (*reproducibility*) e a precisão (*accuracy*) (Quadro 1).

Quadro 1
Tipos de Fiabilidade Relevantes para a Análise de Conteúdo

Tipos de fiabilidade	Design	Erros avaliados	Qualidade relativa
Estabilidade	Teste-reteste	Inconsistências intracodificador	A mais fraca
Reprodutividade	Teste-teste	Inconsistências intracodificador e desacordos entre codificadores	
Precisão	Teste-padrão	Inconsistências intracodificador, desacordos entre codificadores e desvios sistemáticos em relação à norma	A mais forte

Fonte: Adaptado de Krippendorff (1980, p. 131)

A *estabilidade* (também designada de "fiabilidade intracodificador" ou, simplesmente, "consistência") refere-se ao grau de invariabilidade de um processo de codificação ao longo do tempo. Ela diz respeito a situações de 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 ambos os momentos, conclui-se que os resultados são fiáveis. As diferenças verificadas entre a forma como as unidades foram codificadas nos dois momentos podem ser o reflexo de diversos fatores: inconsistências do codificador ("ruído"), mudanças cognitivas que este pode ter experienciado, instruções

de codificação pouco claras, ambiguidades presentes no próprio texto, dificuldades do codificador em interpretar adequadamente as instruções de codificação que lhe foram transmitidas ou meros erros casuais de codificação. Esta é *a forma mais fraca de fiabilidade* e nunca deverá ser utilizada como único indicador da aceitabilidade de uma análise de conteúdo.

A *reprodutividade* (também denominada de "fiabilidade intercodificadores", "acordo intersubjectivo" ou meramente "consenso") designa o grau em que é possível recriar um processo de recodificação em diferentes circunstâncias, com diferentes codificadores. O caso mais típico refere-se à situação de teste-teste, em que dois codificadores aplicam, de forma independente, as mesmas instruções de codificação ao mesmo material, num determinado momento temporal. As diferenças eventualmente verificadas entre as codificações serão o reflexo tanto de inconsistências intracodificadores (acima explicitadas) como de diferenças entre os codificadores (quanto, por exemplo, à forma como interpretam as instruções de codificação), podendo ainda exprimir simples erros aleatórios de codificação.

A *precisão* consiste no grau em que um processo de codificação se conforma funcionalmente com um padrão conhecido. Ela é determinada quando o desempenho de um codificador ou de um instrumento de codificação é comparado com um padrão de desempenho correto conhecido, previamente estabelecido. Esta é a forma mais forte de medição da fiabilidade. Infelizmente, os padrões comparativos que permitiriam o cálculo deste tipo de fiabilidade raramente existem. Não sendo possível, na grande maioria dos casos, optar por ela, a solução mais adequada será recorrer à forma que apresenta a maior qualidade possível: a reprodutividade – os resultados encontrados por um codificador terão de ser, no mínimo, reproduzíveis por outros codificadores, utilizando as mesmas instruções de codificação. É a reprodutividade o tipo de fiabilidade recomendado no presente artigo como patamar mínimo a que qualquer analista de conteúdo deve aspirar.

Uma condição essencial para que exista reprodutibilidade (isto é, fiabilidade intercodificadores) é que os codificadores façam a codificação de forma *independente*. Isto implica que não conversem entre si quanto à codificação a aplicar e que não procurem chegar a consensos prévios sobre que decisões de codificação tomar a respeito de determinados segmentos de texto. Também implica que um codificador que eventualmente ocupe uma posição estatutária mais elevada relativamente aos restantes não use essa posição como fonte de legitimidade para impor o seu próprio entendimento ao(s) outro(s). A existência de comunicação entre os codificadores durante o processo de codificação inflaciona artificialmente o consenso.

A medição do acordo entre codificadores

A medição do acordo entre analistas diferentes é necessária em muitas situações práticas da vida profissional e da investigação. Por exemplo, dois médicos podem ter de avaliar a gravidade de um sintoma num paciente, aplicando uma escala contínua, ou dois psicólogos podem necessitar de aplicar uma escala contínua ou ordinal para classificar o estado de um determinado indivíduo. Nestas situações, será adequado utilizar um índice de acordo de natureza correlacional (Banerjee, Capozzoli, McSweeney, & Sinha, 1999) que meça em que medida a pontuação contínua ou ordinal aplicada por um analista coincide com ou se assemelha à utilizada pelo outro.

Contudo, numa boa parte das situações encontradas pelos investigadores em ciências sociais e humanas, a questão que se coloca é a de determinar o nível de concordância quando se aplicam formas de classificação *nominal* – isto é, que assentam, sobretudo, na emissão de juízos *qualitativos* sobre os dados (por exemplo, decidir se uma determinada afirmação de um entrevistado exprime desencanto com o estilo de direção da sua organização).

A fiabilidade deve exprimir-se num valor quantitativo, que revele *em que grau* os diferentes codificadores estão de acordo quanto à classificação da informação analisada. Este valor quantitativo exprimirá o grau de consonância entre "juízes independentes" quanto ao significado da informação qualitativa em apreço.

No cálculo da fiabilidade de um sistema de categorias, é importante distinguir dois cenários, de complexidade distinta, que se podem colocar aos investigadores:

A) *Análise de conteúdo pré-estruturada*. A informação a analisar é apresentada ao codificador de forma pré-estruturada: o "recorte" já foi feito pelo investigador e o juiz recebe uma grelha que integra uma coleção de unidades de registo, devidamente numeradas. Cabe ao juiz, simplesmente, aplicar a estas unidades de registo os códigos constantes do sistema de categorias, que também lhe terá sido entregue pelo investigador;

B) *Análise de conteúdo aberta*. A informação a codificar é apresentada ao juiz sem qualquer formatação prévia. Por exemplo, o investigador passa-lhe o conjunto integral das transcrições das entrevistas que realizou (ou, normalmente, uma amostra das mesmas), sem qualquer estruturação ou esquematização por via de atos de recorte. Neste caso, cabe ao juiz fazer os próprios recortes, acompanhados dos respetivos atos de codificação, com base no sistema de categorias acima referido.

Por razões de espaço, dar-se-á maior atenção, no presente texto, ao primeiro cenário.

Em qualquer dos casos, é importante sublinhar que, idealmente, para se assegurar uma fiabilidade não enviesada, o investigador responsável pelo estudo *não deve ser incluído* enquanto juiz da informação (Kolbe & Burnett, 1991), podendo, no entanto, treinar adequadamente aqueles que codificarão os dados. Assim sendo, torna-se crucial decidir cuidadosamente quem serão os codificadores. A este respeito, o critério essencial é que sejam pessoas com um nível de instrução semelhante ao do investigador e que tenham a capacidade de compreender a linguagem e os procedimentos que terão de adotar no decurso da análise de conteúdo que irão realizar. Desejavelmente, poderão ser outros investigadores ou alunos envolvidos em processos de pesquisa semelhantes e devidamente orientados, do ponto de vista científico.

A análise de conteúdo pré-estruturada

Neste cenário, o investigador deverá começar por constituir uma lista ou grelha de extratos retirados do total da informação a analisar (*corpus*): isto é, de unidades de registo decorrentes de um processo de "recorte", que serão incluídas, *de forma aleatória*, numa grelha com a seguinte configuração (Quadro 2):

Quadro 2
Estrutura de Grelha de Apresentação das Unidades de Registo para Codificação (Grelha do Investigador)

Nº da Unidade de registo	Unidade de registo	Cod1	Cod2	Acordo (S/N)

Esta grelha compreenderá um mínimo de 10% do total da informação a analisar (isto é, do total de unidades de registo existentes), com um montante nunca inferior a 50 unidades. O investigador deverá também assegurar-se de que nela são inseridas unidades de registo relativas a todas as categorias e subcategorias existentes no seu sistema de categorias.

A grelha apresentada no Quadro 2 será utilizada pelo investigador principal e, no exemplo, exprime uma situação em que existem dois codificadores ("juízes") independentes (Cod1 e Cod2). Existindo mais codificadores, serão acrescentadas mais colunas do lado direito da grelha. Cada codificador receberá uma grelha idêntica,

mas com uma única coluna de codificação (*Cod*), não tendo acesso às decisões de codificação tomadas pelo(s) outro(s) juiz(es).

O investigador entrega a cada codificador um exemplar individual da grelha de codificação, preenchida com as unidades de registo por ele selecionadas para codificação, acompanhado do sistema de categorias a aplicar. Poderá neste momento esclarecer junto de cada juiz a natureza deste sistema e clarificar alguma dúvida que possa surgir.

Posteriormente, de forma independente e com base no sistema de categorias facultado, o codificador atribuirá a cada unidade de registo o código que, a seu ver, melhor lhe corresponda e, finda esta tarefa, devolverá a grelha ao investigador. Este registará então na sua própria grelha (Quadro 2) as codificações realizadas pelos diferentes juízes e assinalará as situações de acordo e de desacordo existentes entre eles. Com base nesta informação, procederá, seguidamente, ao cálculo da taxa de fiabilidade.

Como calcular a fiabilidade?

As primeiras abordagens utilizadas para este propósito basearam-se na proporção (ou percentagem) observada do acordo entre juízes. Este é o processo mais simples e, aliás, o único recomendado pelos autores portugueses que dão atenção ao assunto (ver, por exemplo, os textos de Esteves, 2006, e Vala, 1986). Trata-se, no fundo, de aplicar simplesmente a seguinte fórmula geral:

taxa de fiabilidade = nº de acordos / total de unidades de registo

Esta taxa costuma ser multiplicada por 100, para dar origem a uma percentagem de acordo.

O mesmo cálculo pode, aliás, realizar-se a um nível mais fino, para cada categoria ou subcategoria do sistema de codificação aplicado. Neste caso, para cada categoria/subcategoria e numa situação em que existam dois juízes:

taxa de fiabilidade = 2 (nº acordos) / total codificações 1º juiz + total codificações 2º juiz

Assim, calcula-se primeiro o número de codificações coincidentes entre os dois codificadores. Depois determina-se o total de codificações realizado por cada um na categoria ou subcategoria em causa e soma-se esse valor ao total de codificações realizadas pelo outro juiz na mesma categoria ou subcategoria. Posteriormente, multiplica-se o número de acordos por dois[3] e divide-se esse valor por este último

3 Esta multiplicação é necessária, pois se se somam os totais das codificações realizadas por cada um, também precisamos de considerar duas vezes o número de acordos. Tal multiplicação seria desnecessária se nos limitássemos a dividir o número de acordos pelo número médio de codificações realizadas pelos dois juízes.

total. Como no caso anterior, este valor final pode ser multiplicado por 100 para se obter a percentagem de acordo.

Contudo, embora usada amplamente, a determinação da proporção ou percentagem de acordo não é um método recomendado pela maioria dos especialistas. Estes são praticamente unânimes em afirmar que esta percentagem sobrestima o verdadeiro acordo existente entre os juízes. Cohen (1960) referiu-se-lhe, mesmo, como "a mais primitiva das abordagens" (p. 38).

Na verdade, com base num determinado instrumento de classificação, se duas pessoas codificarem um conjunto de objetos de uma forma completamente aleatória, irão provavelmente coincidir diversas vezes, por mero acaso, nas codificações atribuídas. Krippendorff (1980) admite que este consenso, obtido por acaso, pode abranger até 50% das unidades de registo em análise. O verdadeiro acordo terá então de ser entendido como a confluência das classificações que ocorre *para além* da que se estima que teria acontecido por mero acaso.

Num texto publicado em 1960, Cohen apresentou o teste *kappa* (*k*) como uma forma de medir o acordo entre juízes. O teste foi pensado, inicialmente, para situações em que existem dois codificadores, cada um dos quais classifica, independentemente do outro, *n* sujeitos numa de *m* categorias nominais mutuamente exclusivas e exaustivas. Mais tarde, foi modificado para permitir a sua utilização por múltiplos codificadores (Fleiss, 1971).

A medida de Cohen tem por base a noção, referida anteriormente, de que os casos de acordo observados incluem habitualmente situações em que tal acordo poderá ter ocorrido por mero acaso. Por esta razão, introduz, nos cálculos do grau de acordo, uma correção para a possibilidade de isso ter acontecido. Tecnicamente, isto faz-se confrontando a proporção de acordo observada com o nível de acordo estatisticamente esperado em condições de aleatoriedade das classificações realizadas pelos codificadores envolvidos. Por outras palavras, compara-se o acordo obtido com o acordo que se estima que teria surgido se os codificadores tivessem tomado as suas decisões de uma forma totalmente aleatória.

Segundo Cohen (1960), a probabilidade de o acordo ter sido obtido por acaso é uma função das probabilidades marginais – isto é, da forma específica e relativa como cada codificador distribui os códigos existentes pelas distintas unidades de registo em análise. Baseando-se na lei multiplicativa das probabilidades, este estatístico estimou que a probabilidade de um "acordo por acaso" entre dois juízes independentes consiste no produto das suas probabilidades marginais (independentes) (Perrault & Leigh, 1989).

As proporções esperadas de um acordo por acaso em cada uma das células de uma tabela que cruza as codificações realizadas por dois codificadores são calculadas com base no pressuposto da independência entre os juízes (isto é, de que

cada codificação realizada por um deles não é condicionada pelas dos outros). Este processo é análogo ao utilizado na análise das tabelas de contingência 2x2 com o teste do Qui-quadrado. Vejamos um exemplo (Tabela 1):

Tabela 1
Acordo Hipotético entre Dois Codificadores que Classificam 100 Alunos quanto à Presença ou Ausência de Dificuldades de Aprendizagem

		Codificador 2		
		Tem dificuldades	Não tem dificuldades	Total
Codificador 1	Tem dificuldades	50	15	65
	Não tem dificuldades	15	20	35
	Total	65	35	100

Fonte: Adaptado de Brennan e Silman (1992)

A proporção de acordo observada na Tabela 1 é simplesmente a proporção de diagnósticos coincidentes por parte dos dois codificadores quanto aos alunos que apresentam dificuldades de aprendizagem (50/100 = 0.50), acrescida da proporção de diagnósticos coincidentes quanto à ausência dessas dificuldades (20/100 = 0.20), o que resulta na proporção de 0.70. As proporções esperadas de acordo, atingido por acaso, para cada uma das quatro células da tabela, são calculadas exatamente como se faz para o cálculo do teste do Qui-quadrado. Tendo em consideração que ambos os codificadores classificaram 65% dos alunos como possuindo dificuldades de aprendizagem, então será de esperar que, por simples acaso, os codificadores tenham coincido neste diagnóstico em 65/100 * 65/100 vezes, isto é, em 0.42 (ou 42%) do total de casos analisados. Do mesmo modo, será de esperar que, por mero acaso apenas, tenham coincidido no diagnóstico de que os alunos não possuem dificuldades de aprendizagem em 35/100 * 35/100 casos, isto é, em 0.12 (ou 12%) das ocasiões. A proporção de acordo esperada por mero acaso é, pois, 0.42 + 0.12 = 0.54. O que o teste *kappa* faz é calcular a quantidade extra de acordo observado, depois de se tomar em consideração o acaso, ou o acordo aleatório esperado (*po – pe*), em que *po* é a proporção observada e *pe* a proporção esperada, sobre a quantidade máxima de acordo não aleatório que seria possível ocorrer teoricamente (1-*pe*). Dito de outro modo, o *kappa* exprime a proporção de acordo que não se deve ao acaso, calculada sobre o máximo de acordo não aleatório possível. Portanto, $k = (po-pe)/(1-pe)$. No exemplo, $k = (0.70--0.54)/(1-0.54) = 0.16/0.46 = 0.35$.

Como explica Cohen,

> na medida em que existam fatores não aleatórios a intervir no sentido do acordo, *po* excederá *pe*; a sua diferença, *po-pe*, representa a proporção de casos em que ocorreu acordo para além do acaso e é o numerador do coeficiente. O coeficiente *k* é simplesmente a proporção de desacordos esperados por acaso que não ocorrem ou, alternativamente, a proporção de acordo *depois* de se deixar de tomar em consideração o acordo obtido por acaso (Cohen, 1960, p. 40, itálicos no original).

Daí a fórmula

$$k = (po\text{-}pe)/(1\text{-}pe)$$

ou, exprimida em frequências, para facilitar os cálculos,

$$k = (fo - fe)/(N - fe)$$

Aplicação do *kappa* de Cohen no SPSS

Para aplicar o *kappa* de Cohen no SPSS, deve começar-se por organizar o ficheiro de dados do seguinte modo, no *Data View* do programa (Quadro 3):

Quadro 3
Exemplo da Estrutura do Ficheiro de Dados, no *Data View* do SPSS, para Cálculo do *kappa* de Cohen numa Análise de Conteúdo Pré-Estruturada

Unidade de registo	Codific1	Codific2
1		
2		
3		
...		
...		

Reserva-se assim a primeira coluna para o número de identificação das unidades de registo codificadas, a cada uma das quais o investigador deverá ter atribuído um número de ordem; a segunda coluna é reservada para a inserção dos códigos atribuídos pelo Codificador 1 a estas unidades e a terceira para os códigos atribuídos pelo Codificador 2 às mesmas unidades[4]. Eis um exemplo (Tabela 2) em que dois codificadores atribuíram a 10 unidades de registo os códigos de um sistema de categorias composto pelas categorias 1 (subdividida nas subcategorias 1.1 e 1.2), 2 e 3.

4 A primeira coluna pode ser, aliás, dispensável, se se assumir que cada linha do *Data View* do SPSS representa uma unidade de registo identificada com o respetivo número.

Tabela 2
Exemplo de Inserção dos Dados no SPSS para o Cálculo do *k* de Cohen numa Análise de Conteúdo Pré-Estruturada

Unidade	Codific1	Codific2
1	3.0	3.0
2	1.1	1.2
3	1.1	2.0
4	1.2	1.2
5	1.1	1.1
6	2.0	2.0
7	2.0	1.2
8	3.0	3.0
9	2.0	2.0
10	1.2	1.2

Após a introdução dos dados no formato indicado anteriormente, o investigador procedeu ao cálculo do valor de *kappa*. Para este efeito, seguiu os seguintes passos: nos menus do SPSS, escolheu ANALYZE, DESCRIPTIVE STATISTICS, CROSSTABS. Na caixa de diálogo que se abre, colocou a variável "Codific1" nas colunas e a variável "Codific2" nas linhas (ou vice-versa, pois é indiferente, para efeitos práticos). Seguidamente, pressionou o botão "Statistics" e selecionou a opção "Kappa". Confirmou depois as suas opções com OK.

A partir dos dados da Tabela 2, o programa criou uma tabela que cruza as codificações realizadas por um codificador com as efetuadas pelo outro (Tabela 3). Para além deste resultado, o SPSS também fornece o valor do teste *kappa* e o respetivo nível de significância estatística.

Tabela 3
Resultados da Análise dos Dados da Tabela 2, com a Aplicação do *k* de Cohen no SPSS

		Codific1				
		1.1	1.2	2.0	3.0	Total
Codific2	1.1	1	0	0	0	1
	1.2	1	2	1	0	4
	2.0	1	0	2	0	3
	3.0	0	0	0	2	2
Total		3	2	3	2	10

Nas células situadas na diagonal da tabela (em sombreado) é possível observar o número de vezes em que houve acordo entre os dois juízes. Vemos assim que o Codificador 1 atribuiu o código 1.1 três vezes (total da primeira coluna), enquanto o Codificador 2 o fez apenas uma vez (total da primeira linha), e que ambos aplicaram este código à mesma unidade de registo apenas uma vez (primeira célula do canto superior esquerdo da Tabela 3). O código 1.2 foi aplicado consensualmente duas vezes, acontecendo o mesmo com os códigos 2 e 3. Em suma, em 10 codificações, os juízes concordaram sete vezes.

As frequências das células fora da diagonal indicam não apenas o nível de desacordo, mas também ajudam a perceber a natureza desse desacordo, isto é, em que categorias ele ocorre (Perreault & Leigh, 1989). Vemos, por exemplo, na segunda célula a contar de cima para baixo, na primeira coluna do lado esquerdo, que houve uma unidade a que o Codificador 1 atribuiu o código 1.1, enquanto o Codificador 2 a classificou com o código 1.2. É possível perceber, por exemplo, quantas vezes cada categoria ou subcategoria foi aplicada por cada codificador e em quantas dessas vezes houve coincidência no seu juízo classificativo. A observação desta tabela cruzada pode, pois, ser muito útil para perceber o comportamento dos codificadores e para recolher pistas (por exemplo, que categorias estão a obter consensos muito baixos?) para aperfeiçoar o sistema de classificação, caso isso se venha a mostrar necessário.

O SPSS também produziu, para os dados em apreço, um valor de *kappa* de 0.605 (ou 60.5%), com um nível de probabilidade de $p < 0.01$. Repare-se que aquele valor é inferior à taxa de acordo que teria sido calculada através do método da percentagem (7 acordos em 10 codificações = 70 %). Isto deve-se ao facto de o valor incluir já uma correção que tem em conta a probabilidade de alguns dos acordos se terem devido ao acaso.

Na sua aplicação no SPSS, ao *kappa* surge associado um nível de significância estatística (*p*, representado no programa pela abreviatura *Sig*) cuja interpretação é útil quando trabalhamos com amostras de unidades de registo retiradas de uma "População" maior (ou seja, quando calculamos a fiabilidade com base em apenas algumas das unidades de registo existentes no *corpus*).

Como interpretar os resultados

O *kappa* pode variar entre 1 e -1. O 1 sinaliza um acordo perfeito entre os juízes; o 0 não exprime, como se poderia supor, a ausência de acordo, mas antes a *existência de acordo que se deve totalmente ao acaso* – as codificações idênticas dos juízes são em número igual ao das que teriam acontecido por acaso; o -1 traduz um desacordo perfeito e que não se deve ao acaso (Norusis, 2005, p. 430). Em síntese, o acordo

que supera o esperado (por acaso) conduz a valores positivos; o que é inferior ao esperado dá origem a valores negativos.

Mas como interpretar os valores de *kappa* que se situam no vasto leque de resultados possíveis entre 0 e 1 (em valor absoluto)? Infelizmente, nem todos os investigadores coincidem nos valores de referência para a leitura destes resultados. Por exemplo, Landis e Koch (1977) apresentaram diferentes leques de valores indicativos de distintos níveis de acordo. Embora os próprios autores tenham designado os seus valores de "claramente arbitrários", estes acabaram por ser adotados na literatura como o padrão de referência para a interpretação do *kappa*. De acordo com estas indicações (também assumidas por Fleiss, 1981), valores superiores a 0.75 sinalizam um acordo forte que está para além do acaso, valores inferiores a 0.40 representam um baixo nível de acordo para além do acaso, e valores entre 0.40 e 0.75 representam um acordo de razoável a bom, não obtido por acaso.

Alternativamente, em Brennan e Silman (1992), encontramos a seguinte tabela de referência (Tabela 4):

Tabela 4
Interpretação Sugerida por Brennan e Silman (1992) para os Diferentes Valores de *kappa*

Kappa	***Grau de acordo***
<0.20	Fraco
0.21-0.40	Razoável
0.41-0.60	Moderado
0.61-0.80	Bom
0.81-1.00	Muito Bom

Dada esta diversidade de orientações, recomenda-se que o analista opte por um modelo interpretativo específico e que o cite explicitamente, para que seja possível ao leitor perceber em que quadro intelectual e técnico se situa a sua análise.

Uma condição importante para que seja possível calcular o *k* no SPSS é que ambos os codificadores usem as mesmas categorias e subcategorias, isto é, torna-se necessário que o programa possa construir uma tabela de dupla entrada *simétrica* em que as categorias sejam idênticas nas linhas e nas colunas. Mais concretamente, não pode dar-se o caso de haver uma categoria que só um dos codificadores usou, mas não o outro.

Esta é uma limitação do *kappa*. Com efeito, existem dois tipos gerais de desacordo possível entre codificadores: (a) um pode entender que na unidade de registo 1 se deve aplicar o código x, enquanto o outro acha que deve ser aplicado o código y; (b) um pode considerar que um determinado código é aplicável a, pelo menos, uma unidade de registo, enquanto o outro considera que tal código não se aplica a nenhuma. O *kappa* de Cohen só está construído para lidar com a primeira destas situações e

o SPSS dá mensagem de erro quando se confronta com o segundo contexto. Mas, se virmos bem, este não é propriamente um problema: é muito provável que um sistema de categorias no qual algumas categorias nunca sejam aplicadas por um ou mais codificadores tenha problemas intrínsecos e a mensagem de erro do SPSS funcionará até, a este respeito, como um sinal de alarme de que é necessário aperfeiçoar o sistema (ou, hipótese que também é admissível, o treino dos codificadores).

O que fazer quando a fiabilidade é baixa?

Suponhamos que no seu estudo um investigador conclui que a taxa de acordo entre os juízes é baixa. O que deve fazer? Nestas situações, deverá procurar aperfeiçoar o seu sistema de categorias e repetir todo o processo de codificação e determinação da taxa de fiabilidade (preferencialmente, com outros juízes, ou outras unidades de registo), até atingir um limiar de fiabilidade adequado. Nesta fase, antes de repetir todo o processo, tendo em vista melhorar o sistema de categorias, recomenda-se que converse com os codificadores e procure perceber que aspetos acharam difíceis na aplicação do sistema. Como resultado deste diálogo e da própria análise que o investigador faz dos comportamentos de codificação dos juízes (como se referiu, anteriormente), pode ser necessário aplicar uma ou mais das seguintes medidas:

a) proceder à fusão de categorias e/ou subcategorias;

b) alterar os descritivos das categorias e/ou subcategorias, para tornar mais clara a sua natureza;

c) inserir exemplos típicos mais adequados para ilustrar o conteúdo indicativo dessas categorias e/ou subcategorias.

A análise do *corpus*

Depois de se atingir o nível de fiabilidade desejado, aplica-se o sistema de categorias a *toda* a informação disponível. Cada conjunto de unidades de registo colocadas numa determinada categoria ou subcategoria poderá, então, ser analisado, de forma a compreender-se em que sentido aponta a informação nele contida. Para o efeito, devem organizar-se ficheiros temáticos (por exemplo, no processador de texto WORD), um para cada categoria ou subcategoria do sistema aplicado. Em alternativa, poderá utilizar-se um programa informático de análise qualitativa, de entre os vários disponíveis no mercado. O conteúdo de cada ficheiro será posteriormente analisado, procurando-se detetar as principais tendências de resposta,

comportamento ou perceção encontradas e relacionar as conclusões decorrentes da análise dos diferentes ficheiros.

Embora alguns autores utilizem a contagem de excertos ou de palavras/expressões como estratégia analítica, em cada ficheiro temático, este não é decididamente o critério de análise mais importante, embora possa constituir um auxiliar útil na apresentação e interpretação dos resultados. O mais relevante é a componente qualitativa da informação, embora até os partidários mais puristas deste tipo de análise não consigam normalmente resistir a fazer alguma espécie de contagem numérica para decidir o que é importante nos seus dados.

Conclusão

Não existe uma forma *certa* de se fazer análise de conteúdo e o presente trabalho não pretende apresentar-se como a solução perfeita para todas as dificuldades com que este tipo de trabalho se confronta. Procurou-se, mesmo assim, dar um passo em frente, fazendo um balanço crítico sobre as formas habituais como temos lidado com estas situações e sugerindo processos mais estruturados e explícitos de as aperfeiçoarmos. Para que um analista possa verificar em que medida cumpriu os requisitos enunciados no presente texto, apresenta-se em anexo uma lista de verificação (Anexo B).

Neste artigo, deu-se especial atenção à questão da fiabilidade, por se considerar que é um elemento central nos esforços que precisamos de desenvolver para conferirmos às nossas análises de conteúdo um estatuto mais rigoroso no âmbito da comunidade científica. Na sua variante "intracodificador", o termo fiabilidade descreve em que medida um mesmo codificador codifica o material para análise de uma forma consistente, em momentos diferentes. Procurámos demonstrar que, apesar de importante, esta forma de fiabilidade é insuficiente para assegurar a credibilidade científica de um estudo, sendo essencial a busca da fiabilidade "intercodificadores". Enquanto a primeira mede simplesmente a consistência dos juízos privados de um sujeito, a segunda estabelece um grau de consistência baseado em entendimentos *partilhados*, isto é, assente na confluência de juízos formulados por dois ou mais codificadores.

Mas é preciso reconhecer que a obtenção de um acordo entre juízes quanto à codificação de uma unidade de registo ou de um segmento de texto não é garantia de que essa unidade ou segmento tenha sido codificada corretamente: ambos os juízes podem ter errado na atribuição do código "correto". Por isto, no presente texto insistiu-se na importância de se medir a fiabilidade com base numa forma de cálculo

que não exprima uma simples percentagem de acordo, sugerindo-se, para o efeito, a utilização do *kappa* de Cohen.

Embora o *kappa* tenha algumas limitações, que foram aqui assinaladas, é importante usá-lo, porque (a) é a medida mais utilizada para se calcular a fiabilidade intercodificadores no domínio das ciências comportamentais, o que permite a comparação entre estudos; (b) constitui a base sobre a qual a grande maioria dos outros testes e abordagens foi construída e (c) toma em consideração a probabilidade de uma parte dos acordos obtidos se ter devido ao acaso (Perrault & Leigh, 1989).

Contudo, o *kappa* não é uma solução milagrosa para as complexidades inerentes ao processo de cálculo da fiabilidade na análise de conteúdo. O risco de a taxa de fiabilidade ser baixa aumenta com a quantidade de categorias a aplicar, o número de codificadores que intervêm e o grau de desestruturação do material a codificar. Para se obter maior fiabilidade, "basta", portanto, reduzir o número de categorias a aplicar e o número de codificadores a mobilizar, apresentando-lhes, preferencialmente, unidades de registo previamente recortadas pelo investigador. Isto mostra que existe uma certa *artificialidade metodológica* no cálculo da fiabilidade que não deve ser menosprezada. O contributo essencial dos procedimentos recomendados no presente texto reside, pois, não propriamente na descoberta da forma "ideal" de se calcular a fiabilidade – embora a técnica sugerida represente um avanço em relação ao que tem sido habitual fazer-se entre nós –, mas, sobretudo, em dois outros aspetos: (1) o estabelecimento de uma forma tecnicamente clara de se organizar a informação e o processo de análise e de se assegurar um acordo intersubjetivo entre analistas e (2) a insistência na explicitação e revelação pública dos procedimentos e das decisões tomadas, permitindo assim que a análise de conteúdo se torne verificável e, portanto, digna de um estatuto científico.

Referências bibliográficas

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond Kappa: a review of interrater agreement measures. *The Canadian Journal of Statistics / La Revue Canadiènne de Statistique, 27*(1), 3-23.

Bardin, L. (1995). *Análise de conteúdo*. Lisboa: Edições 70.

Brennan, P., & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures. *British Medical Journal, 304*, 1491-1494.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Constas, M. A. (1992). Qualitative analysis as a public event: the documentation of category development procedures. *American Educational Research Journal, 29*(2), 253-266.

Esteves, M. (2006). Análise de conteúdo. In J. Á. Lima & J. A. Pacheco (Orgs.), *Fazer investigação* (pp. 105-126). Porto: Porto Editora.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2ª edição). New York: John Wiley and Sons.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8-18.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: an examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*(2), 243-250.

Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Newbury Park, CA: Sage.

Landis, R. J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: assessment and reporting of intercoding reliability. *Human Communication Research, 28*(4), 587-604.

Norusis, M. J. (2005). *SPSS 13.0 Guide to data analysis*. New Jersey: Prentice-Hall.

Perreault, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research, 36*, 135-148.

Rocha, D., & Deusdará, B. (2005). Análise de conteúdo e análise do discurso: aproximações e afastamentos na(re)construção de uma trajetória. *Alea, 7*(2), 306-322.

Vala, J. (1986). A análise de conteúdo. In A. S. Silva & J. M. Pinto (Orgs.), *Metodologia das ciências sociais* (pp. 101-128). Porto: Afrontamento.

Weber, R. (1990). *Basic content analysis*. Newbury Park, CA: Sage.

Anexo A

Exemplo de um sistema de categorias para análise de conteúdo

Categoria 1 - DESENVOLVIMENTO PROFISSIONAL INDIVIDUAL

Esta categoria agrega os excertos das entrevistas relativos ao conceito de desenvolvimento profissional manifestado pelos entrevistados, às experiências pessoais de desenvolvimento profissional que relatam, aos obstáculos a esse desenvolvimento que consideram existir na sua escola e às condições que, na sua opinião, seriam necessárias para que se desenvolvessem mais, desse ponto de vista.

Subcategoria 1.1 - Conceito de desenvolvimento profissional

Engloba as referências produzidas pelos entrevistados que exprimem o conceito que possuem de desenvolvimento profissional no ensino. Integra a sua descrição das mudanças e experiências verificadas ao longo do tempo na sua prática e nas suas atitudes que indiciam um desenvolvimento *individual*, enquanto docente.

Exemplo:

"Desenvolver-me, enquanto docente, é sentir que estou a crescer interiormente e que consigo comunicar e fazer com que os alunos cresçam, também, enquanto pessoas."

Subcategoria 1.2 - Obstáculos ao DP individual existentes na escola

Integra as afirmações dos entrevistados que ilustram a sua opinião sobre que aspetos existem na sua escola que constituem obstáculo ao seu desenvolvimento profissional *individual*.

Exemplo:

"Podia-me ter desenvolvido bastante mais se tivesse tido mais apoio. Apoio de quem percebesse realmente de educação especial."

Subcategoria 1.3 – Condições promotoras do DP individual na escola

Integra as afirmações dos entrevistados que ilustram a sua opinião sobre que aspetos existem na sua escola que promovem o seu desenvolvimento profissional *individual*.

Exemplo:

"Tendo em conta os recursos que nós temos na escola ... é uma escola na qual eu vejo que há ritmo, que há dinamização, que há tentativa de ... de disponibilizar formação aos professores de modo a melhorarmos."

Subcategoria 1.4 - Condições necessárias para um maior DP individual

Agrupa as referências feitas pelos entrevistados às condições que seriam necessárias para que se desenvolvessem mais, *individualmente*, do ponto de vista profissional.

Exemplo:

"Dava jeito, por exemplo, sair mais vezes daqui, que é uma coisa que nós fazemos muito pouco. (...) gostava que houvesse um bocadinho mais de intercâmbios e de experiências, trocas ... partilha de experiências."

Categoria 2 – DESENVOLVIMENTO DA ESCOLA

São incluídas nesta categoria as referências feitas pelos inquiridos às condições que seriam necessárias para que a sua escola se desenvolvesse mais, enquanto estabelecimento de ensino.

Exemplos:

"Eu acho que [o que a escola precisa para se desenvolver mais] é o aspeto ... eu acho ... o aspeto humano, acho que falha um bocado aí. Eu acho que tem que se ter em conta mais a pessoa, o lado humano dos docentes, ver também as suas ... os seus problemas que fazem com que às vezes tenham um desempenho menos ... menos bom."

"A escola precisa de adquirir mais material ... estou sempre a lembrar-me da minha área, mais material de laboratório, material informático, se calhar talvez permitir mais consulta por parte dos alunos, só temos duas salas de Internet ligadas em rede."

Anexo B

Análise de Conteúdo: Lista de Verificação dos Procedimentos Metodológicos

Na minha análise de conteúdo:	Sim	Não
1. Explicitei o modo como construí o sistema de categorias, identificando as categorias pré-determinadas e as emergentes		
2. No meu sistema de categorias, não existem unidades de registo enquadráveis em mais do que uma categoria ou subcategoria		
3. As categorias e subcategorias do meu sistema de categorias são exaustivas, isto é, exprimem todos os aspetos relevantes existentes nos dados		
4. O meu sistema de categorias é produtivo: todas as categorias/subcategorias têm pelo menos três unidades de registo		

5. Usei um sistema decimal de atribuição de códigos às categorias/subcategorias		
6. No sistema de categorias, incluí em cada categoria e subcategoria uma descrição por extenso da sua natureza		
7. No sistema de categorias, incluí em cada categoria e subcategoria um ou dois exemplos típicos de unidades de registo enquadráveis na mesma		
8. Escolhi codificadores com nível de instrução e capacidade de compreensão linguística e técnica adequados		
9. Realizei um treino adequado dos codificadores, explicando-lhes detalhadamente o sistema de categorias		
10. Utilizei pelo menos dois codificadores		
11. Não intervim no papel de codificador		
12. Os codificadores agiram de forma independente		
13. No cálculo da taxa de fiabilidade, utilizei uma técnica que tem em conta a possibilidade de o acordo poder ocorrer por acaso		
14. Para o cálculo da fiabilidade, usei uma amostra aleatória de pelo menos 10% das unidades de registo, com um mínimo de 50 unidades*		
15. Apresentei o nível de significância estatística da taxa de fiabilidade apurada*		
16. Alterei o sistema de categorias e procedi a novo cálculo da fiabilidade, quando o cálculo anterior produziu uma taxa baixa**		

*** critério aplicável unicamente quando a fiabilidade é calculada com base numa amostra de unidades de registo ou de segmentos de texto**

**** critério aplicável apenas nas situações em que a taxa calculada anteriormente se revela insatisfatória**

A Case for a More Reliable Content Analysis

Abstract

Content analysis is a technique that is very often referred to and used in the social and human sciences, namely in educational research, but it is rarely performed in an adequate way. This paper argues for a more systematic and rigorous approach to this type of analysis and proposes technical solutions for some of the main situations with which content analysts confront themselves. This paper includes methodological recommendations with respect to several aspects of this kind of work and includes a checklist that allows researchers to determine the extent to which their analyses meet criteria of rigour in this domain.

Key-words: content analysis; reliability; Cohen's *kappa*

Vers une Analyse de Conténu Plus Fiable

Résumé

L'analyse de contenu est une technique très invoquée et utilisée dans les sciences sociales et humaines, incluant les sciences de l'éducation, mais elle est rarement développée de façon correcte. Dans ce texte, on défende la nécessitée d'une approche plus systématique et rigoureuse a cette analyse et on propose des solutions techniques pour quelques situations principaux que les analystes de contenu rencontrent. Le texte inclut recommandations méthodologiques relatives a divers aspects de ce type de travail et une liste de vérification qui permet aux chercheurs déterminer en quelle mesure ses analyses se conforment aux critères de rigueur en cet domaine.

Mots-clés: analyse de contenu; fiabilité; *kappa* de Cohen