# ENTRE A LIBERDADE DE INVESTIGAÇÃO E A PROTEÇÃO DE DADOS SENSÍVEIS. O CASO DO *CORPUS* CORAUDIS

## *Between research freedom and sensitive data protection: the case of the CorAuDis corpus*

CONCEIÇÃO CARAPINHA
*mccarapinha@fl.uc.pt*
*Universidade de Coimbra, CELGA-ILTEC, Faculdade de Letras*

CORNELIA PLAG
*cornelia.plag@fl.uc.pt*
*Universidade de Coimbra, CELGA-ILTEC, Faculdade de Letras*

SUSANA FERREIRA
*susana.ferreira@fl.uc.pt*
*Universidade de Coimbra, CELGA-ILTEC, Faculdade de Letras*

ORCID: https://orcid.org/0000-0001-7860-6561
https://orcid.org/0000-0002-5644-8723
https://orcid.org/0000-0004-1633-431x

DOI
*https://doi.org/10.14195/0870-4112_3-11_10*

Texto recebido em / Text submitted on: 14/04/2025
Texto aprovado em / Text approved on: 04/09/2025

**Biblos**. Número 11, 2025 • 3.ª Série
pp. 237-259

**RESUMO**

A pesquisa baseada em *corpora* representa uma forma de validar a investigação; contudo, a disponibilização dos textos em regime de acesso aberto pode levantar questões relacionadas com a partilha de dados pessoais dos envolvidos. No caso de um *corpus* de audiências, este problema é substancialmente ampliado.

O presente artigo aborda os desafios encontrados no processo de anonimização do *corpus* de audiências CorAuDis, bem como as questões éticas suscitadas pelos dados sensíveis nele contidos. São também analisadas as possíveis formas de o disponibilizar em regime de acesso aberto. Este contexto suscita ainda uma reflexão sobre a liberdade científica do investigador, que procura encontrar um equilíbrio entre os requisitos de confidencialidade e o valor científico dos textos.

**Palavras-chave:** *Corpus* de audiências; anonimização; dados sensíveis; liberdade científica; acesso aberto.

**ABSTRACT**

*Corpus*-based research is a method of validating investigations; nevertheless, the provision of texts through open access may give rise to questions concerning the sharing of personal data of those involved. This problem is exacerbated in the case of a *corpus* of hearings.

The present article addresses the challenges encountered in the process of anonymising the *corpus* of hearings CorAuDis, as well as the ethical issues raised by the sensitive data it contains. The potential avenues for the dissemination of the aforementioned material in an open access environment are also analysed. This context also rises questions regarding the scientific freedom of researchers, seeking to strike a balance between confidentiality requirements and the scientific value of the texts.

**Keywords:** *Corpus* of hearings; anonymisation; sensitive data; scientific freedom; open access.

## INTRODUÇÃO

No atual paradigma científico, é exigido ao linguista a pesquisa a partir de dados autênticos.

Por conter dados efetivamente produzidos por falantes reais em contextos específicos, por se encontrar informatizado, permitindo, portanto, uma pesquisa mais célere e mais seletiva, e ainda pelo facto de constituir um manancial de informações que podem fundamentar pesquisa linguística em diferentes planos de descrição, um *corpus* pode tornar-se "a useful tool for discovering many aspects of language use that otherwise may go unnoticed" (Reppen, 2022: 13).

Neste sentido, um dos principais objetivos que presidem à construção de um *corpus* é a possibilidade de o disponibilizar como infraestrutura de apoio à investigação, especialmente entre pares. Também a crescente adoção de práticas científicas abertas vai ao encontro deste desiderato; com a Declaração de Bona (2020), reafirma-se a promoção da liberdade de investigação científica no espaço europeu e promove-se "o direito dos investigadores individuais de definir livremente as questões que são objeto de investigação, (...) de reunir dados empíricos, (...) de publicar e comunicar livremente e de propor novas ideias e hipóteses e de as divulgar livremente".

Em sentido divergente ao estabelecido nestes princípios, todavia, emergem, na construção de um *corpus*, quer na fase da recolha de dados, quer no momento do seu tratamento informático, quer ainda na etapa da sua disponibilização, múltiplas e complexas questões que nem sempre são fáceis de compaginar com os princípios consignados na Declaração. Uma dessas questões diz respeito à confidencialidade dos dados a compilar e ao grau de liberdade do investigador.

Se muitos dos *corpora* atualmente disponíveis não envolvem problemas quanto à divulgação dos dados neles contidos, o mesmo não se verifica quando os dados a compilar são de natureza sensível e/ou carreiam informação personalizada. O *corpus* em construção que deu origem à presente reflexão – o *corpus* de audiências CorAuDis – constitui um destes casos, exigindo uma reflexão sobre o direito à proteção de dados pessoais dos envolvidos e sobre o dever de confidencialidade que pode recair sobre o investigador que os colige e analisa.

Tendo como enquadramento estas questões, o presente texto apresenta e discute os desafios encontrados – quer no plano mais prático quer no plano mais teórico – aquando do processo de anonimização das transcrições, refletindo ainda sobre o papel do investigador, dividido entre as exigências científicas e o direito à liberdade de investigar, por um lado, e as responsabilidades éticas, por outro.

As perguntas que orientaram esta pesquisa foram as seguintes:

– Em que consiste o processo de anonimização[1]?

– Quais os problemas práticos enfrentados pelos investigadores no tratamento de informação sensível?

– Qual pode/deve ser o grau de intervenção do investigador?

Este artigo apresenta a seguinte organização: na secção 1, abordar-se-ão alguns dos princípios que orientam a construção de um *corpus* e os problemas gerais encontrados durante a transcrição das audiências; a secção 2 discutirá as diferentes modalidades de anonimização segundo o Regulamento Geral de Proteção de Dados (RGPD)[2] e os desafios práticos e teóricos decorrentes do processo de anonimização do *corpus* CorAuDis; uma reflexão sobre a condição e o papel do investigador no atual panorama da ciência aberta e eticamente responsável ocupará a secção 3. O texto encerrará com algumas considerações relativas aos moldes em que o *corpus* poderá ser disponibilizado.

## 1. A CONSTRUÇÃO DE *CORPORA*

### 1.1. PRINCÍPIOS DE COMPILAÇÃO DE DADOS

Os *corpora* podem ser bastante distintos, quer no tocante à quantidade de dados recolhida, quer no que respeita ao seu tipo, à sua diversidade, à sua acessibilidade,

[1] O termo 'anonimização' é usado, aqui, num sentido genérico. Ver, adiante, a secção 2.

[2] A norma que assegura a execução, na ordem jurídica nacional, do Regulamento (UE) 2016/679 do Parlamento e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares e à livre circulação desses dados (RGPD) é a Lei n.º 58/2019 de 08 de agosto.

quer ainda no que concerne à forma de os organizar internamente; contudo, há alguns princípios que devem orientar a compilação de um *corpus*.

Os textos autênticos devem ser representativos dos textos da área em análise, ou seja, conter uma amostra significativa do universo em estudo, permitindo, assim, efetuar generalizações. Um *corpus* deve também possuir uma outra propriedade, a do equilíbrio, o que implica apresentar uma distribuição proporcional dos diferentes textos abrangidos pelo domínio em causa (neste caso, a 'audiência', enquanto evento comunicativo). O conjunto de dados que constituem o *corpus* deve ainda estar armazenado em formato digital, viabilizando pesquisas automáticas ou semiautomáticas através do recurso a *software* especializado. Um outro princípio diz respeito à autenticidade ou, por outras palavras, à preferência pela preservação da totalidade dos textos: segundo Sinclair (1991), a remoção de informação de um *corpus* deve ser limitada, para evitar a perda de dados e de autenticidade[3]. É precisamente à luz deste princípio que será discutida a anonimização de dados, uma vez que a remoção de informação relevante pode reduzir ou até anular o potencial valor científico dos dados[4].

### 1.2. DESAFIOS DA TRANSCRIÇÃO DO *CORPUS* CORAUDIS

O CorAuDis constitui uma infraestrutura de apoio à investigação, concebida com o propósito de possibilitar o estudo sistemático de um género discursivo oral, produzido num contexto profissional e institucional. Trata-se de um *corpus* especializado que permite descrever regularidades discursivas e padrões de uso linguístico próprios deste *setting*. É constituído por um conjunto de 37 audiências, da área do Direito Penal (relativas aos anos de 2016 e 2017), que teve como base os áudios disponibilizados pelo Tribunal de Coimbra[5]. Na impossibilidade –

[3] Um conjunto mais completo de princípios de construção de *corpora* consta de Knight e Adolphs (2022: 23).

[4] A mesma ideia é retomada, embora de forma mitigada, em Knight e Adolphs (2022: 23).

[5] A gravação das audiências que o próprio Tribunal realiza constitui um procedimento habitual, relacionado com o recurso que pode advir das decisões judiciais.

legal – de divulgar os ficheiros áudio, deu-se início ao processo de transcrição, o único formato permitido para a divulgação dos dados.

A transcrição de uma audiência, gravada apenas através de material áudio, é um processo moroso e exigente (Love e Wright, 2021). A multiplicidade de intervenientes dificulta a identificação precisa das vozes e, portanto, dos autores dos enunciados. O ruído de fundo e a baixa qualidade das gravações do Tribunal comprometem, frequentemente, a audibilidade e inteligibilidade do discurso, duas condições essenciais para garantir a precisão e o rigor da transcrição.

Emergem, ainda, com particular relevância, os problemas decorrentes do registo oral em que a audiência decorre. A representação de dados paralinguísticos (pausas, silêncios, tom de voz, débito, etc.), de fenómenos fonéticos (de uma pronúncia mais marcada ou de crases e aféreses, por exemplo) e de traços discursivos (hesitações, repetições, interrupções, formas truncadas, e.o.) exigiu um minucioso sistema de transcrição. Tendo em conta estas especificidades, é inegável que algumas perdas resultarão do processo de transcrição. E, apesar dos avanços significativos da tecnologia e da crescente adoção de ferramentas automáticas, como "voice-to-text software and (...) AI rather than human transcription" (McMullin, 2021: 140), parece não haver nenhum sistema de transcrição que consiga, de forma rigorosa e exaustiva, dar conta de toda a complexidade e especificidade do oral.

A opção por uma transcrição *full verbatim* – transcrição literal – pretendeu preservar toda a informação possível, apesar da perda de dados de natureza contextual, decorrente da não utilização de vídeo (na gravação original) e da perda de outros dados, motivada pelo próprio processo de transcrição. Paralelamente, procurou-se assegurar a inteligibilidade do texto, evitando o uso excessivo de sinalética.

Particularmente complexos foram os dilemas colocados pelo processo de anonimização de dados pessoais, dado que uma transcrição *ipsis verbis* das gravações teria permitido a identificação de pessoas, locais e eventos mencionados nas sessões. Tornou-se óbvio que seria impossível transcrever a totalidade do que havia sido dito em sala de audiências, uma vez que questões de ordem ética e legal impendiam sobre o conteúdo dos áudios. Desta

forma, o já referido princípio da preservação da totalidade dos textos teve de ser repensado, pois, na verdade, "[w]hile an ideal scenario would be to include all possible data produced in a given context in order to claim, in a scientific manner, a total accountability, this is in practice rarely possible" (Jaworska e Kinloch, 2018: 7).

Como proceder, então, para não divulgar dados pessoais e, simultaneamente, tentar manter o máximo possível de informação acerca do caso que conduziu à audiência? Em rigor, toda a discussão em torno deste problema conduziu às três perguntas de base (ver introdução).

## 2. O PROCESSO DE ANONIMIZAÇÃO

### 2.1. ENQUADRAMENTO LEGAL

A anonimização é o processo que remove as ligações criadas entre uma pessoa específica e os seus dados (Quinn, 2017), impossibilitando o respetivo rastreamento (Stam e Diaz, 2023)[6].

Os dados pessoais englobam diferentes tipos de informações sobre as pessoas; segundo as normas previstas no RGPD, entende-se por "dados pessoais" toda a

> informação relativa a uma pessoa singular identificada ou identificável («titular dos dados»); é considerada identificável uma pessoa singular que possa ser identificada, direta ou indiretamente, em especial por referência a um identificador, como por exemplo um nome, um número de identificação, dados de localização, identificadores por via eletrónica ou a um ou mais elementos específicos da identidade física, fisiológica, genética, mental, económica, cultural ou social dessa pessoa singular; (...).

[6] Este método de desassociação de dados pessoais constitui, aliás, uma obrigação legal, prevista pelo Parlamento Europeu, aplicada à ciência e aos organismos públicos e privados.

Neste vasto conjunto de informações, Leedham et al. (2021: 2) distinguem identificadores diretos e indiretos: os diretos correspondem a informações individuais e únicas que permitem, por si só, identificar uma pessoa (Stam e Diaz, 2023: 8; Leedham et al., 2021: 2), como, por exemplo, o nome completo, o número do cartão de cidadão ou a voz; no caso dos indiretos, a identificação é possível através do cruzamento de dados. Embora a idade, por si só, não constitua um elemento identificador, a sua combinação com outros dados, tais como o cargo profissional ou o local de trabalho, por exemplo, pode conduzir à pessoa em causa[7].

De acordo com o Considerando 51 do RGPD, alguns destes dados podem ser classificados como especialmente sensíveis – a origem étnica, por exemplo –, embora o documento refira também que quaisquer dados pessoais podem, através de um uso indevido, comprometer os direitos e as liberdades fundamentais do seu titular, hipótese que justifica a inclusão destas informações no conjunto dos dados sensíveis.

De modo a minimizar tanto quanto possível a exibição de dados pessoais, o RGPD prevê duas modalidades de encriptação: a anonimização e a pseudonimização.

Enquanto a anonimização é irreversível, pois todos os dados pessoais são removidos, de forma permanente, tornando-se irrecuperáveis, a pseudonimização é o processo de substituição de informação sensível por informação neutra, reduzindo as hipóteses de identificação, a não ser recorrendo a informações adicionais, ou seja, ao conjunto de metadados que se manterão fora do alcance do público[8]. Com efeito, esta segunda técnica permite a conservação das tais "informações suplementares", como disposto no Artigo 4.º, alínea 5), do Regulamento:

[7] A este propósito, Stam e Diaz (2023) distinguem identificadores indiretos fortes e fracos, de acordo com a maior ou menor probabilidade de permitirem identificar um indivíduo: veja-se a diferença entre um número de telemóvel e a data de nascimento.

[8] Este arquivo de metadados encontra-se sob a proteção do Artigo 12.º da Lei n.º 59/2019 da República Portuguesa.

> 5) «Pseudonimização», o tratamento de dados pessoais de forma que deixem de poder ser atribuídos a um titular de dados específico sem recorrer a informações suplementares (...) mantidas separadamente e sujeitas a medidas técnicas e organizativas para assegurar que os dados pessoais não possam ser atribuídos a uma pessoa singular identificada ou identificável; (...).

À medida que o grau de personalização dos dados aumenta, mais complexo e moroso se tornará o processo de os anonimizar, exigindo "technical skills and an understanding of privacy requirements to select the most adequate techniques and settings to provide the best anonymization results" (Ferreira et al., 2024: 457)

Uma vez substituídos, os dados adquirem um carácter genérico, tendencialmente inócuo, tendo como principal função preencher espaços em branco, sem comprometer o sentido das transcrições dos dados originais e o trabalho inferencial do leitor/investigador que pretende aceder aos dados (como ilustra o Exemplo 1).

> Exemplo 1[9]:
> J (00.21) – O nome dos pais:: ↑
> A1 (00.22) – eh:: NOME APELIDO2 e NOME APELIDO3.
> J (00.26) – A data de nascimento ↑
> A1 (00.28) – DATA.
> […]
> J (00.32) – =DATA. {ruídos} *Ond*' é que nasceu?
> A1 (00.34) – eh:: Em CONCELHO1.

Em ambos os processos de encriptação, é necessário garantir que os identificadores são transformados ou removidos, ao mesmo tempo que é necessário assegurar a conservação do valor de base do *corpus*, ou seja, garantir que essas alterações não impedem as pesquisas para as quais o *corpus* foi constituído. E este duplo objetivo nem sempre parece ser concretizável.

[9] As abreviaturas presentes nos exemplos correspondem a: J = juiz; A1 = arguido 1; DA4 = defesa do arguido 4; T1 = testemunha 1.

### 2.2. O PROCESSO DE ANONIMIZAÇÃO NO *CORPUS* CORAUDIS

No caso do *corpus* CorAuDis, o processo de anonimização ocorreu em simultâneo com o processo de transcrição.

Numa primeira fase, procedeu-se ao levantamento dos identificadores diretos, que se reportam, no caso em análise, ao número de processo do tribunal, ao nome completo dos envolvidos e, em contextos criminais muito específicos, como o tráfico de drogas e o roubo, às alcunhas e aos hipocorísticos, bem como dos indiretos mais recorrentes, tais como datas, endereços, contactos, características de veículos, locais, empresas, etc., com base na classificação de Stam e Diaz (2023). Numa segunda fase, e por meio da pseudonimização, esses identificadores foram convertidos em elementos vazios demitidos de um titular, "replacing personal information with standard placeholders" (Petyko et al., 2022: 13), tal como se ilustra na tabela *infra*:

| Identificadores (diretos e indiretos) | Após pseudonimização |
|---|---|
| Processo de tribunal | NÚMERO DE PROCESSO |
| Nomes | NOME APELIDO1, NOME APELIDO2, ... |
| Alcunhas[10] | ALCUNHA NOME1, ALCUNHA NOME2, ... |
| Hipocorísticos | HIPOCORÍSTICO NOME1, HIPOCORÍSTICO NOME2, ... |
| Datas | DATA |
| Endereços | CONCELHO, RUA, NÚMERO, CÓDIGO POSTAL, … |
| Contactos | NÚMERO DE TELEMÓVEL |
| Veículos | MARCA, MODELO, MATRÍCULA |
| Locais | HOSPITAL, ESCOLA, ESQUADRA, EMPRESA, … |

[10] Foi decidido que as alcunhas e os hipocorísticos deveriam ser representadas conforme o Exemplo 2. Assim, a um nome próprio específico (*e.g.* Francisco Silva) corresponde uma alcunha (*e.g.* Mãozinhas) ou um hipocorístico (*e.g.* Chico) pela/pelo qual um indivíduo é nomeado dentro da sua rede de amigos ou de contactos.

Exemplo 2:
J (02.47) – (...) permanecendo nas imediações da residência do senhor NOME APELIDO2 (.), também p- > conhecido *pla* alcunha, ALCUNHA NOME2, (…)

Com o avanço do processo de transcrição e a emergência de novos dados, essa listagem de identificadores foi sendo ampliada, exigindo a permanente atualização das normas de transcrição.

Ainda que as substituições acima mencionadas não ofereçam dúvidas e não constituam um obstáculo ao entendimento do evento comunicativo em si mesmo, uma vez que o processo de pseudonimização ocorre num plano muito micro, o mesmo não ocorre quando outro tipo de dados é equacionado.

A reconstituição de acontecimentos passados que tem lugar em sala de audiências surge sob forma de narrativas (muitas vezes opostas) mediadas subjetivamente por quem as viveu ou testemunhou (Silva, 2012). Esse relato de uma história, que se constrói a várias vozes, fica registado na gravação e, *a posteriori*, na transcrição. E, do mesmo modo que algumas destas narrativas subjacentes aos processos judiciais são conhecidas do grande público, por envolverem pessoas vulneráveis ou figuras públicas, as histórias de vida menos mediáticas não deixam de ser importantes, traumáticas e íntimas para quem as experienciou, merecendo igual reserva. Como se pretende demonstrar, não é suficiente empreender a anonimização de nomes (tomados aqui num sentido muito genérico), pois todo o conteúdo de uma audiência é suficientemente claro para permitir a identificação dos envolvidos e da história a eles associada (Rock, 2001). A exposição pública de um destes casos, que pode ocorrer com a disponibilização, sem restrições, de um *corpus* deste género, é, pois, problemática, implicando desafios mais complexos. De facto, é este tipo de "highly sensitive or disturbing data, which poses an important ethical dilemma for researchers" (Petyko et al., 2022: 11).

O presente exemplo reporta-se a uma audiência que julga um crime de furto de energia elétrica, alegadamente cometido pelos proprietários de um estabelecimento comercial numa localidade pequena, contra uma empresa fornecedora de eletricidade. Para além da óbvia presença dos nomes de todas estas entidades, são mencionados diferentes detalhes sobre a vida dos envolvidos, tais como o facto de o estabelecimento ser familiar, isto é, ser propriedade de um casal idoso e do seu filho, o facto de haver um familiar menor a viver com o pai e com a mãe em semanas alternadas, o facto de este pai ter uma casa própria que está a pagar ao banco e ter uma viatura

com um conjunto de características específicas, adequadas ao tipo de negócio desenvolvido e ainda o facto de o referido estabelecimento ter sofrido obras. Apesar de todos os identificadores diretos e indiretos (nos termos do RGPD) terem sido pseudonimizados, há, como se constata, um conjunto de informações disseminadas ao longo da audiência que, quando combinadas, possibilitaria a identificação dos três arguidos.

O mesmo tipo de risco pode ocorrer com o discurso de outros intervenientes, neste caso, dos próprios profissionais. Ainda que contra todas as expectativas, as intervenções destes agentes judiciais têm suscitado muitas interrogações e grandes dificuldades ao processo de ocultação de dados pessoais, por conterem comentários que veiculam visões de mundo muito personalizadas, aumentando assim as possibilidades de serem identificadas como pertencendo a um indivíduo particular. Analisemos, em primeiro lugar, o discurso dos advogados.

Com a exposição de versões antagónicas por parte das duas fações em litígio, a atividade verbal que decorre em tribunal é claramente conflitual. Assim, é expectável que os advogados adotem um tom mais hostil e, naturalmente, mais emocional. Por outro lado, se aos leigos não é permitido desviarem-se dos tópicos propostos pelos profissionais, o discurso destes, pelo contrário, está repleto de digressões, as quais constituem terreno fértil para a partilha espontânea de informações e opiniões pessoais. Toda esta partilha de dados é suscetível de criar dificuldades ao investigador/transcritor, que terá de tomar decisões sobre o que pode ser incluído e o que deve ser omitido.

Numa das audiências, no momento das alegações finais, um dos advogados discorre, nos seus frequentes comentários laterais, sobre a sua atividade profissional e as suas relações – ao que os dados indicam, bastante conflituosas – com os restantes agentes judiciários. As expressões que usa para referir outros membros do seu grupo profissional (Exemplo 3) e as afirmações que faz a propósito do Ministério Público (Exemplo 4) constituem um bom exemplo destes casos, os quais facilmente permitem chegar ao seu nome após uma breve pesquisa.

Exemplo 3:
DA4 (00.00) – (…) Foi feita uma boa corrupta!

Exemplo 4:
DA4 (00.00) – (...) É por isso *qu'*a justiça, em Portugal, (.) passa *plo* desprestígio que passa! (.) É DEVIDO ao Ministério Público, {ruídos} ao MAU USO que este faz do processo, no inquérito e no julgamento, *QU'* AS COISAS CHEGARAM AO PONTO A QUE CHEGARAM! (.) Não tenho (problemas) {ruídos} de o dizer, *porqu'*eu DOU A CARA E:: DOU NA CARA! PORTANTO, NÃO MANDO RECADOS A NINGUÉM![11]

Tal exposição pública de dados pessoais (opiniões e crenças) dificulta enormemente a tarefa de transcrição. A omissão desta informação, no sentido de evitar a identificação do jurista, implicaria a eliminação quase total de grande parte do final da audiência, o que, por sua vez, comprometeria a relevância científica do texto.

A este propósito, e embora o RGPD categorize estes identificadores indiretos como "especialmente sensíveis" e determine que o seu tratamento é proibido, abre algumas exceções "para fins de arquivo de interesse público, para fins de investigação científica ou histórica ou para fins estatísticos baseados no direito da União ou dos Estados-Membros" (Considerando 53). Entramos, então, numa zona cinzenta, pois o Regulamento, que rotula estes dados de sensíveis, abre, em simultâneo, a possibilidade de os analisar (e expor), ao abrigo do pressuposto da chamada Ciência Aberta.

Um outro exemplo, de natureza diferente, ilustra ainda um movimento argumentativo baseado em estereótipos que categorizam comunidades e os seus comportamentos:

Exemplo 5:
DA1 (15.11) – Olhe! eh:: São pessoas de etnia cigana?
T1 (15.14) – Sim.
DA1 (15.15) – eh:: Geralmente, estas situações, *quand'* um é detido, (.) a família junta-se toda, não é? (.) É costume?

[11] As maiúsculas representam um tom de voz mais forte e/ou mais irritado.

Há ainda a assinalar, no âmbito do discurso dos profissionais, o discurso dos magistrados, aquele que tem originado mais ponderação.

A longa e exigente formação de um magistrado permitiria antecipar que o conhecimento legal e a experiência teriam, como consequência, um desempenho imparcial em sala de audiências. De facto, o artigo 34.º da Lei n.º 2/2008, que regula o ingresso nas magistraturas e a formação de magistrados, bem como o Plano de Estudos do 41.º Curso de Formação de Magistrados para os Tribunais Judiciais (2024/2025), do Centro de Estudos Judiciários, preconiza, entre outras "competências fundamentais inerentes ao exercício de funções", a de "alcançar a melhor solução do ponto de vista prático e jurídico (justiça do caso concreto), no âmbito de um procedimento que respeite, no plano substantivo e processual, os direitos de todas as pessoas envolvidas".

Porém, no material recolhido para o *corpus*, sobressai todo o tipo de enviesamentos e de estereótipos que perpassam no discurso dos juízes. A imparcialidade e a neutralidade, perante quem litiga e perante o objeto do litígio, qualidades essenciais na prática judicial (Almeida, 2017), encontram-se, por vezes, bastante afastadas do fórum, sendo substituídas por juízos de valor subjetivos.

Se muitos destes enviesamentos afloram de modo implícito, não consciente, podendo, ainda assim, afetar a forma como o juiz avalia os factos e a forma como delibera, pois, de acordo com Harris e Sen (2019: 242), "research shows that judges' personal backgrounds, professional experiences, life experiences, and partisan and ideological loyalties might impact their decision making", muitos outros podem, de forma mais saliente, ser diretamente observados – e gravados! – quando são verbalizados de forma explícita pelos profissionais.

Os vários exemplos subsequentes revelam que muitas perguntas e muitos comentários do juiz contêm observações que, à luz de uma análise mais crítica, não se coadunam com as exigências éticas previstas na sua formação. Veja-se o caso seguinte, relativo à inquirição de uma testemunha (vítima) de um crime de ameaça agravada:

> Exemplo 6:
> J (16.42) – Ó senhor NOME APELIDO4! O senhor > O senhor acha esse seu comportamento normal? (.) Se calhar, aqui ninguém na sala acha! (..)

> O senhor está a exponenciar um facto *qu'* a generalidade das pessoas, apesar de ter algum temor, não entra em DEPRESSÃO! (.) Está a perceber o > Está a perceber a[sic] porquê > o porquê das perguntas da *sotora*? {ruídos} Tem a ver essencialmente *co'* isso! O senhor é *duma* sensibilidade EXTREMA!

Na mesma audiência, o juiz inquire, depois, o filho da vítima da ameaça. Atente-se, mais particularmente, na primeira e na última intervenções do magistrado e na forma como é designada a autora do crime de ameaça, também presente em sala de audiência.

> Exemplo 7:
> J (06.06) – Um rapaz de dezoito anos [tem medo] > *duma* > *duma* velha de > de > de::/
> T1 (06.10) – Ah! Isso:: > Isso quer dizer muito pouco!
> J (06.13) – (( )) que tem receio o quê? *Qu'* ela tenha lá uma caçadeira em casa?
> T1 (06.16) – {riso} Isso não sei.
> J (06.18) – Ou *qu'* ela ponha uma bomba e mande aquilo tudo *plos* ares?
> DA? (06.21) – Ó meu Deus! [(( )) {riso}
> J (06.21) – [Não! Estou > Agora estou > Agora estou a ques- > estou a questionar! Precisamente porque não percebo *com'* é *qu'* um rapaz de dezoito anos, numa situação destas, também > também se d- > se > se:: > se fica na sombra dos medos do pai!

Numa outra audiência, o juiz – que dialoga com os restantes profissionais acerca de um dos participantes leigos ali presentes – recorre ao uso da terceira pessoa gramatical para o referir, ignorando a sua presença e negando, ao visado, o estatuto de verdadeira pessoa. Em simultâneo, a utilização do termo 'diminuído', para além de cientificamente discutível, também não promove um ambiente inclusivo.

> Exemplo 8:
> J (05.14) – [\a dizer, efetivamente, *s'* ele é capaz ou incapaz, {conversas

> paralelas} *s'* é diminuído ou não é diminuído, *pra* eles não terem problemas. Por isso, no dia três *d'* abril, o senhor virá cá! É *memo pa* alegações e::, depois::, marcamos o:: > a leitura do acórdão. *Tá* bem?

No exemplo 9, é utilizado o termo 'raça', que também é um termo cientificamente equivocado.

> Exemplo 9:
> J2 (10.10) – Os *outos* senhores que *tavam* na linha > na linha de reconhecimento eram da me- > eram de raça negra?

No caso seguinte, e ainda que em tom jocoso, é bem patente a reprodução de ideias estereotipadas acerca de uma determinada comunidade:

> Exemplo 10:
> J (05.34) – *Sotor*, ela é russa, mas não havia indícios de vodca!

Em suma, o discurso do juiz nem sempre é neutro e pode revelar a adoção de visões de mundo e atitudes tendenciosas, que correspondem à *vox populi*, e que ele reproduz acriticamente em sala de audiência.

## 3. O INVESTIGADOR: ENTRE A LIBERDADE CIENTÍFICA E A RESPONSABILIDADE ÉTICA

Partilhar dados, tornando os *corpora* públicos, constitui uma salutar prática académica, recomendada pelas universidades e pelas entidades financiadoras da investigação. O direito à liberdade de pesquisa e de publicitação dos dados e dos resultados obtidos parece, contudo, neste caso particular, sofrer algumas constrições, dada a especificidade e o caráter sensível do material que compõe o *corpus*. O domínio do discurso forense – tal como o da saúde – exige, aos investigadores, uma reflexão aprofundada sobre os riscos envolvidos na divulgação de toda a informação recolhida em sala de audiências.

Em muitos casos, a investigação contendo dados pessoais pode prosseguir sem limitações quando se obtém o consentimento informado dos participantes. Porém, no *corpus* em análise, e considerada a origem das audiências, obtidas através de uma fonte privada (Nelson, 2010: 61), que mantém as gravações apenas para uso interno da instituição, não é possível aceder aos produtores originais dos discursos. Sem conseguir obter o consentimento livre e informado dos diferentes participantes, resta, ao investigador, o recurso à anonimização (ou pseudonimização) processo que, como se constatou, não é isento de problemas e de aporias.

Ainda que o material áudio coligido corresponda a audiências abertas ao público, é lícito considerar o objetivo – meramente técnico-jurídico – para o qual o Tribunal obtém as gravações; em rigor, elas são usadas apenas em caso de recurso para tribunais superiores, os quais as avaliam no sentido de confirmar se os factos ficaram provados e/ ou se a lei foi corretamente aplicada, sem qualquer outra finalidade (considerando que cada juiz é soberano na sua audiência). Em sentido divergente, os objetivos da análise que pode ser realizada através da disponibilização do *corpus* em acesso aberto são completamente distintos, bastante mais amplos, envolvendo, pelo menos potencialmente, a interseção com outras áreas científicas (a Linguística, em primeiro lugar, mas também a Sociologia, a Antropologia, a Ciência Política, o próprio Direito...) que podem ter uma visão mais crítica dos fenómenos ali observados. De facto, os discursos dos agentes judiciários, caracterizados por enviesamentos de vária ordem, poderiam, se amplamente divulgados, dar origem a estudos sobre a ideologia dos juízes, os preconceitos e os estereótipos veiculados.

Terá o investigador a responsabilidade social de trazer a público e de questionar estes discursos, considerando, por outro lado, a inexistência de uma autorização formal – de cada um dos magistrados envolvidos – para a sua divulgação? Terá o investigador legitimidade para expor e, no fundo, denunciar estes enviesamentos quando os objetivos originais do projeto – apresentados ao Tribunal – não eram estes?

Parece haver, em suma, dois objetivos irreconciliáveis por parte do investigador: por um lado, reivindicar o direito à liberdade científica, tentando

contribuir para a melhoria da administração da justiça – demonstrando, por exemplo, através de dados reais, a forma como a comunicação entre profissionais e leigos pode não funcionar, a forma como a linguagem usada em tribunal pode constituir um obstáculo no acesso do cidadão comum à Justiça ou ainda a forma, discriminatória, como algumas comunidades são tratadas em sala de audiências; por outro lado, cumprir a legislação relativa à proteção de dados.

Os programas de Ciência Aberta promovem a disponibilização dos dados resultantes dos projetos de investigação em plataformas de Acesso Aberto; no entanto, essa abertura é sempre condicionada pelos requisitos legais relativos à privacidade dos indivíduos envolvidos na investigação, o que reconduz a presente reflexão ao processo de anonimização.

Em busca de respostas para os dilemas surgidos, sobretudo para o processamento destes dados considerados sensíveis, foram testados alguns programas informáticos que poderiam agilizar o fluxo de anonimização/pseudonimização para dados escritos. No entanto, nenhum deles[12] reconhece os segmentos textuais mais problemáticos, os que contêm marcas de enviesamento[13]; em paralelo, e tal como vimos, o RGPD também é taxativo quanto à necessidade de evitar a publicitação de dados sensíveis:

> (75) O risco para os direitos e liberdades das pessoas singulares, cuja probabilidade e gravidade podem ser variáveis, poderá resultar de operações de tratamento de dados pessoais suscetíveis de causar danos físicos, ma-

[12] Os programas (Argus, ARX, sdcMICRO, CAT, UTD, entre outros) testados no âmbito do projeto CorAuDis não se revelaram um coadjuvante na tarefa de anonimização de dados sensíveis, já que todos eles exigem conhecimentos avançados de informática. Até à data, nem sequer o trabalho colaborativo entre algumas destas aplicações e a IA favorece o reconhecimento de segmentos textuais que encerram marcas de discurso tendencioso.

[13] Tendo em conta a complexidade das aplicações testadas, foi solicitada a demonstração de um *software* pago – *Data Anonymizer* –, uma aplicação *online*, com uma interface intuitiva, que substitui, automaticamente, os identificadores diretos e indiretos por valores genéricos, de acordo com as configurações estabelecidas pelo utilizador. Ainda que seja um investimento que permita economizar tempo, o problema persiste: a máquina não está treinada para detetar os discursos enviesados apresentados neste estudo.

> teriais ou imateriais, em especial quando o tratamento possa dar origem à discriminação, à usurpação ou roubo da identidade, a perdas financeiras, prejuízos para a reputação (...).

A primeira conclusão a tirar é a de que os dados forenses, como estes que constituem o CorAuDis, não podem ser tornados públicos tal como foram recebidos, sob pena de poderem acarretar danos reputacionais irreversíveis para os envolvidos (que nem sequer autorizaram a sua divulgação); em segundo lugar, conclui-se também que os atuais programas de anonimização de dados em uso em diferentes áreas (inclusive no contexto judicial) não estão adequadamente desenhados para responder aos distintos desafios colocados por este material.

O que fazer, então? Qual o grau de liberdade científica do investigador, que tem de fazer uso de procedimentos de pseudonimização e, tem, portanto, de intervir e de manipular os dados, mantendo, em simultâneo, a sua utilidade, não diminuindo a inteligibilidade das transcrições, e não comprometendo a partilha de conhecimento científico? Quais os limites entre divulgação e transgressão?

A adoção de um único modelo de ética na ciência será o mais adequado? Exigirão os dados qualitativos com que se trabalha nesta área um outro modelo? Resumir-se-á o processo de anonimização à identificação dos itens lexicais que devem ser encriptados ou, como defende Krishmanurthy (*apud* Rock, 2001), deve ser o conteúdo, na sua totalidade, a merecer a atenção do investigador? E se esta for a resposta, os procedimentos de anonimização terão de ser manuais e a análise, casuística. Ainda assim, e tal como questiona Baker (2018), será possível remover, completamente, toda a informação pessoal dos *corpora*?

## CONSIDERAÇÕES FINAIS

No final da presente reflexão, necessariamente limitada pelo número de audiências transcritas (e pelos problemas nelas encontrados), é pertinente realçar alguns pontos.

Os dados que mais dúvidas e hesitações desencadearam e que motivaram a presente reflexão – disseminados no discurso dos advogados e dos magistrados – localizam-se num plano bem mais impercetível do que aquele em que se encontram os identificadores diretos, ou seja, situam-se no plano da expressão da subjetividade por parte dos profissionais do fórum, por outras palavras, no plano da expressão de opiniões, crenças e visões de mundo. Por outro lado, e de forma distinta da dos identificadores diretos, a codificação destes traços estende-se a diferentes planos de descrição linguística para além do lexical (fonológico, morfológico, sintático, semântico e discursivo) (Pinto et al., 2021: 207) tornando, neste caso, mais complexo ainda o processo de identificar esses traços e encriptá-los.

Perante a dificuldade em encontrar estudos que respondam, de forma concludente, a estas questões, perante a ausência de legislação que defina, de forma inequívoca, os procedimentos a seguir, ou até de práticas consolidadas em outras áreas de ciência, a decisão final sobre as medidas a adotar quanto às modalidades de disponibilização do *corpus* só pode ser cautelar.

Num contexto em que os conceitos de anonimização, confidencialidade e acesso aberto se intersetam, o estabelecimento de filtros de acesso ao *corpus* emerge, então, como um procedimento urgente a implementar. Se o conhecimento do discurso que ocorre em sala de audiências é essencial para realizar estudos, para melhorar a comunicação (ou a falta dela) em tribunal e, no fundo, para fazer serviço público, consolidando valores, como a 'justiça social', é, todavia, fulcral enfatizar que as restrições de confidencialidade e o respeito pelos dados pessoais se sobrepõem a todas as aspirações trazidas pelo modelo da Ciência Aberta.

Prevê-se, assim, que as versões originais das audiências sejam arquivadas num repositório detido pela instituição que tutela a investigação e que apenas estejam disponíveis para os investigadores afetos ao projeto; para a versão pseudonimizada dos textos – que manterão os identificadores indiretos que foram discutidos na secção 2.2. –, o acesso será limitado aos investigadores que satisfizerem um conjunto de critérios: será necessário o registo, acompanhado da explicitação dos objetivos subjacentes à pesquisa que pretendem efetuar e, mais importante ainda, um comprovativo de que integram um centro de investigação.

Alguns problemas se antecipam, ainda assim, no horizonte. Um deles reporta-se aos custos financeiros e logísticos inerentes ao processo de implementação destes filtros, dada a necessidade de monitorização.

Um outro possível problema diz respeito ao uso que os investigadores qualificados farão dos dados. Ainda que com todas as credenciais validadas, não será possível (nem eticamente justificável) controlar o tipo de estudos e de divulgação de dados que virá a ser realizado por esses investigadores.

Um terceiro aspeto, já de natureza distinta, que terá de ser equacionado, talvez de forma longitudinal, é o dos efeitos da anonimização nos estudos que vierem a ser realizados, dada a possibilidade de ser necessário refinar, ainda, os critérios de anonimização, à medida que novas audiências vierem a ser transcritas[14].

Não existindo uma resposta única para todas as questões levantadas, o presente estudo, de natureza qualitativa, pode ajudar a fundamentar ou a definir futuros protocolos que permitam melhorar os procedimentos de transcrição de dados forenses ou, até, a construção de *corpora* de dados sensíveis.

## BIBLIOGRAFIA

Almeida, Maria Teresa F. (2017). Julgar com uma perspetiva de género? *Julgar*, 1-13. https://julgar.pt/wp-content/uploads/2017/11/20171109-ARTIGO-JULGAR-Julgar-com-uma-perspetiva-de-g%C3%A9nero-Teresa-F%C3%A9ria.pdf.

Assembleia da República (2008). Lei n.º 2/2008. Diário da República n.º 9/2008, Série I de 2008-01-14. https://files.diariodarepublica.pt/1s/2008/01/00900/0039100412.pdf.

— (2019). Lei n.º 58/2019. Diário da República n.º 151/2019, Série I de 2019-08-08. https://files.diariodarepublica.pt/gratuitos/1s/2019/08/15100.pdf.

— (2019). Lei n.º 59/2019. Diário da República n.º 151/2019, Série I de 2019-08-08. https://files.diariodarepublica.pt/gratuitos/1s/2019/08/15100.pdf.

Baker, Paul (2018). Language, sexuality and corpus linguistics. concerns and future directions. *Journal of Language and Sexuality*, 7, 2, 263-279.

[14] À data de publicação deste artigo, o CorAuDis tem 15 transcrições completas, estando a 16.ª em curso.

Centro de Estudos Judiciários (2024). Plano de Estudos do 41.º Curso de Formação de Magistrados para os Tribunais Judiciais (2024/2025). *Ed. Centro de Estudos Judiciários*. https://cej.justica.gov.pt/Portals/30/Ficheiros/formacao/inicial/curso_41/1.%C2%BA_ciclo/Plano_de_Estudos_41.%C2%BACurso.pdf?ver=a2SNIu6WJibD1S1VjFklbw%3d%3d.

Ferreira, Ana; Bischoff, Francisco; Almeida, Rute; Nogueira-Silva, Luis; Cruz-Correia, Ricardo; Muchagata, Joana (2024). How Anonymous Are Your Anonymized Data? The AnyMApp Case Study. *Communications in Computer and Information Science*, vol 1958. Cham: Springer. https://doi.org/10.1007/978-3-031-49215-0_54.

Harris, Allison P.; Sen, Maya (2019). Bias and Judging. *Annual Review of Political Science*, 22, 1, 241-259. http://dx.doi.org/10.1146/annurev-polisci-051617-090650.

Jaworska, Sylvia; Kinloch, Karen (2018). Using multiple data sets. In C. Taylor and A. Marchi (Eds.), *Corpus Approaches to Discourse: A Critical Review* (110-129). London: Routledge.

Knight, Dawn; Adolphs, Svenja (2022). Building a spoken corpus: what are the basics? In Anne O'Keeffe and Michael McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (21-34). London: Routledge.

Leedham, Maria; Lillis, Theresa; Twiner, Alison (2021). Creating a corpus of sensitive and hard-to-access texts: Methodological challenges and ethical concerns in the building of the WiSP corpus. *Journal of Applied Corpus Linguistics*, 1, 3, 1-10. https://doi.org/10.1016/j.acorp.2021.100011.

Love, Robbie; Wright, David (2021). Specifying Challenges in Transcribing Covert Recordings: Implications for Forensic Transcription. *Frontiers in Communication*, 6. https://doi.org/10.3389/fcomm.2021.797448.

McMullin, Caitlin (2021). Transcription and Qualitative Methods: Implications for Third Sector Research. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 34, 1, 140-153. https://link.springer.com/article/10.1007/s11266-021-00400-3.

Nelson, Mike (2010). Building a written corpus: what are the basics? In Anne O'Keeffe and Michael McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (53-65). London: Routledge.

Parlamento Europeu. (2024). Resolução do Parlamento Europeu P9_TA(2024)0022. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0022_EN.html.

Petyko, Marton; Busso, Lucia; Grant, Tim; Atkins, Sarah (2022). The Aston Forensic Linguistic Databank (FoLD). *Language and Law / Linguagem e Direito*, 9, 1, 9-24.

Pinto, Alexandra Guedes; Warrot, Catarina Vaz; Cardoso, Henrique Lopes; Duarte, Isabel Margarida; Sousa-Silva, Rui (2021). Deteção de linguagem tendenciosa em decisões judiciais. *Revista da Associação Portuguesa de Linguística*, 8, 203-217.

Quinn, Paul (2017). The Anonymisation of Research Data – A Pyric Victory for Privacy that Should Not Be Pushed Too Hard by the EU Data Protection Framework?. *European Journal of Health Law*, 24, 1-21.

Reppen, Randi (2022). Building a corpus: what are key considerations? In Anne O'Keeffe and Michael McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (13-20). London: Routledge.

Rock, Frances (2001). Policy and Practice in the Anonymisation of Linguistic Data. *International Journal of Corpus Linguistics*, 6, 1, 1-26.

Silva, Joana Aguiar e (2012). As narrativas do direito e a verdade judicial. In Rui do Carmo (Coord.), *Linguagem, argumentação e decisão judiciária* (111-124). Coimbra: Coimbra Editora.

Sinclair, John (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Stam, Alexandra e Diaz, Pablo (2023). Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts. *FORS Guides*, 20, Version 1.1, 1-15. https://doi.org/10.24449/FG-2023-00020.

União Europeia (2016). Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, relativo à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados. https://eur-lex.europa.eu/eli/reg/2016/679/oj.