FALA-EM-INTERAÇÃO MEDIADA POR VÍDEO: DESAFIOS NA TRANSCRIÇÃO DE GESTOS FACIAIS

Ulrike Schröder[*]
Flavia Fidelis de Paula[**]

▪ RESUMO: A pandemia restringiu as interações presenciais, enquanto acelerou a globalização das conexões humanas por meio de videochamadas. Isso possibilitou a realização em larga escala de atividades como trabalho colaborativo, conferências, ensino, consultas médicas, ritos religiosos e encontros familiares. Essas práticas não apenas promoveram a expansão das fronteiras locais, mas também conectaram o mundo globalmente, o que fortaleceu encontros interculturais caracterizados cada vez mais por conversas multilíngues. Nesta nova configuração, as "cabeças falantes" (*talking heads*) assumem uma importância central. O presente trabalho apresenta e compara dois modelos metodológicos para a transcrição de gestos faciais em interações mediadas por vídeo, conforme propostos por Brunner e Diemer (2021) e Dix (2022). Escolhemos estas duas propostas atuais dentro de uma discussão sobre transcrição refinada no campo da Análise da Conversa Multimodal, como uma primeira tentativa de testar a aplicabilidade desses modelos em interações mediadas por vídeo. Nosso estudo ressalta a necessidade de desenvolver metodologias que capturem a complexidade multimodal dos gestos faciais nessas interações, ao mesmo tempo que demonstra como esses sistemas podem contribuir para métodos de pesquisa quantitativos e qualitativos que integram dimensões linguísticas e gestuais. Além disso, discutiremos as vantagens e desvantagens dos dois sistemas e faremos propostas para seu aperfeiçoamento.

▪ PALAVRAS-CHAVE: Transcrição multimodal; Interação mediada por vídeo (IMV); Inglês como língua franca, Comunicação intercultural.

Introdução

A ascensão do mundo digital-virtual já estava em curso quando a pandemia de COVID-19 acelerou a transição para uma nova normalidade. Essa nova realidade nos

[*] Universidade Federal de Minas Gerais (UFMG), Faculdade de Letras, Belo Horizonte, MG, Brasil. Professora Titular. schroederulrike@gmx.com. ORCID: https://orcid.org/0000-0001-7764-7249

[**] Universidade Federal de Minas Gerais (UFMG), Faculdade de Letras, Belo Horizonte, MG, Brasil. Pesquisadora. flaviafid@gmail.com. ORCID: https://orcid.org/0000-0002-9427-8745

impeliu rapidamente a aprender a interagir por meio de diversas tecnologias de mediação por vídeo e transcendeu fronteiras do local para o global. As pessoas passaram a utilizar videoconferências para conversar com amigos, colaborar no trabalho, participar de reuniões, consultar profissionais de saúde, participar de congressos internacionais, expandir atividades institucionais e engajar-se em rituais religiosos, bem como ampliar suas experiências de lazer por meio do mundo virtual. Sendo assim, as restrições à interação presencial impostas pela pandemia foram acompanhadas por uma rápida abertura de diversas comunidades multilíngues e interculturais também.

No que diz respeito a esse universo digital-virtual, antes da pandemia, as pesquisas na área de linguística, comunicação e interação estavam predominantemente centradas em dados de redes sociais e plataformas como Facebook, WhatsApp, X (antigo Twitter) ou YouTube. A atenção à fala-em-interação virtual é um tópico mais emergente, notadamente na Análise da Conversa Multimodal (Due; Licoppe, 2021) e na Educação a Distância (EaD) (Salomão; Freire Junior, 2021). Em uma "interação virtual" *online* e síncrona os interlocutores se conectam por meio de tecnologias que proporcionam acesso simultâneo ao som e à imagem em tempo real (Stivers; Timmermans, 2017; Fidelis de Paula, 2023, p. 12). Por reconhecemos a necessidade de explorar essa dinâmica de maneira linguística, interacional e multimodal, em 2022, iniciamos o projeto de pesquisa *The Multimodal Coordination of Intercultural Video-Mediated Interaction*.[1] Este projeto visa investigar a interação mediada por vídeo sob uma abordagem multissensorial e corporificada *in situ*, além de explorar a crescente dimensão (inter)cultural da interação *online*. Uma pergunta-chave que se coloca perante a nova epistemologia da interação virtual é a questão da transcrição. A Análise da Conversa (AC), a Linguística Interacional (LI) e os Estudos de Gestos (EdG) influenciaram significativamente outros campos dos Estudos Linguísticos durante as últimas décadas, o que contribuiu para o entendimento um "turno corporificado" ou "turno multimodal" (Mondada, 019; Neville, 2015) nas investigações da língua em uso. Porém, a transição para a interação virtual levanta questões essenciais, especialmente no que diz respeito à transcrição. O corpo, antes central nas investigações linguísticas, passa por uma certa transposição invertida, dada a predominância de "cabeças falantes" (*talking heads*, Licoppe; Morel, 2012) na comunicação virtual. Esta mudança levanta questões sobre a atenção dos participantes, os dados fornecidos pela tela e pela filmagem, e, destarte, os desafios específicos para uma transcrição adequada.

Sendo assim, a necessidade de refletir sobre a transcrição de interações virtuais se deve principalmente a dois aspectos: (i) o imparável uso do espaço digital para a comunicação humana, potencializado desde a pandemia e (ii) uma lacuna metodológica que se baseia, entre outros, em uma certa negligência dos gestos faciais que é acompanhada por uma integração insatisfatória desses gestos nos sistemas de transcrição disponíveis. No Referencial Metodológico, abordaremos essa questão a partir de uma discussão das características específicas da interação virtual. Em seguida, discutiremos

[1] www.letras.ufmg.br/icmi

os avanços e as lacunas na transcrição multimodal, com foco especial nos gestos faciais. Propomos uma classificação de gestos faciais e suas funções com base em Bavelas e Chovil (Bavelas, 2022), alinhada aos pilares teóricos da AC, da LI e dos EdG. Na fundamentação metodológica, apresentaremos dois sistemas atuais desenvolvidos para enfrentar desafios na transcrição multimodal: (i) GAT 2+ISWA (Dix, 2022), um sistema orientado ao "turno visual", e (ii) ViMELF (Brunner; Diemer, 2021), projetado para lidar com dados ricos e *corpora* de Inglês como Língua Franca (ILF) mediados por vídeo. Ilustraremos o funcionamento desses sistemas com um exemplo curto e concluiremos com uma discussão das vantagens e desvantagens.

Referencial metodológico

Interação virtual no mundo global

A investigação da fala-em-interação virtual em comparação com a abundância de outros fenômenos no mundo digital ainda está em estado embrionário. Durante as últimas duas décadas, ocorreu uma "virada corporificada" no campo de pesquisa da interação, cognição e linguagem, que iniciou uma dissociação do dualismo neo-cartesiano em direção a uma nova reivindicação de intercorporeidade (Streeck, 2021).[2] De acordo com essa visão, (inter)ação, gesto, linguagem e cognição estão interconectados e são fenômenos inerentemente sociais, uma vez que os indivíduos habitam naturalmente as ações dos outros. Portanto, uma abordagem praxeológica e fenomenológica à interação *in situ* é baseada em uma compreensão holística da comunicação, que também concebe gestos como "acoplados ao ambiente"[3] (Goodwin, 2018). Além dos recursos corporais-visuais, gestos manuais e, especialmente no caso de interações *online* e síncronas, gestos faciais (Kendon, 2004; Bavelas; Gerwing; Healing, 2014), é a prosódia que assume uma função crucial na interação incorporada. Sob a influência da LI, a prosódia ganhou um novo foco de pesquisa, abrangendo "os atributos "musicais" da fala – efeitos auditivos como melodia, dinâmica, ritmo, tempo e pausa"[4] (Couper-Kuhlen; Selting, 1996, p. 11).

Desde o início de suas atividades em 2010, nosso Centro de Pesquisa *Intercultural Communication in Multimodal Interactions* – ICMI[5] tem como base empírica interações interpessoais e interculturais não mediadas, ou seja, a presença física era concebida como um conceito de espaço compartilhado, de intimidade e de imediatismo. Em contraste, a

[2] Adotamos, neste artigo, uma perspectiva rigorosamente comunicativa, ou seja, *from within*, em conformidade com o paradigma da Análise da Conversa, fundamentado na fenomenologia social de Alfred Schütz e em estudos no campo da etnografia. Discussões relacionadas a outros campos de pesquisa ultrapassam o escopo deste artigo.

[3] No original: "environmentally coupled".

[4] No original: "the 'musical' attributes of speech – auditory effects such as melody, dynamics, rhythm, tempo and pause".

[5] www.letras.ufmg.br/icmi

interação mediada por vídeo implica uma série de mudanças em relação aos encontros presenciais. Essas diferenças são de alta relevância em termos epistemológicos, metodológicos e analíticos e devem ser consideradas ao abordar dados provenientes de interações mediadas por vídeo. Em resumo (cf., entre outros, Due; Licoppe, 2020; Friesen, 2014; Loenhoff, 2012; Licoppe; Morel, 2012; Manstead; Lea; Goh, 2011; Norris; Prini, 2016; Balaman; Doehler, 2022):

(a) em vez de interações de corpo inteiro, os interlocutores interagem em um "aglomerado ecológico olho a olho" (*talking heads*);
(b) os espaços visual e auditivo são fragmentados (iluminação, cenografia, segmentação corporal, características vocais, ruído de fundo etc.);
(c) ocorrem interrupções visuais (resolução da tela) e auditivas (*staccatos*, ecos, volume etc.);
(d) não há espaços de transição para comunicação informal (elevadores, corredores etc.) em casos de interações institucionais;
(e) o modo básico de experiência é 2D em vez de 3D (por exemplo, quando objetos entram em jogo);
(f) a distância interpessoal é geralmente reduzida (e.g., em caso de reuniões, palestras etc.);
(g) os interlocutores precisam lidar com o dilema do contato visual, uma vez que os movimentos dos olhos não orientam a atenção, a alocação e a tomada de turnos etc.;
(h) a atenção é dividida, pois os interlocutores controlam também sua própria imagem transmitida.

Diante desses fenômenos importantes que já foram revelados como tendo um impacto crucial na interação *online* e síncrona, deve-se mencionar que eles podem se tornar ainda mais relevantes na comunicação multilíngue e intercultural, situações em que a prevenção de mal-entendidos (*pre-empting strategies*), por exemplo, torna-se altamente relevante.

Recentemente, estudiosos da AC seguiram um caminho em direção a uma metodologia de coleta de dados baseada em vídeo, com o objetivo de explorar, em um nível muito detalhado e granular, como a interação mediada por vídeo é realizada momento a momento. Ao fazer isso, tornou-se evidente que existem muitas perguntas em aberto, antigas e novas, que surgem devido a esse novo contexto, relacionadas à transcrição. Na próxima seção, chamaremos a atenção para essa questão.

As armadilhas da transcrição multimodal

A transcrição "multimodal" não é tão recente como parece. Sua raiz está na pesquisa etnográfica, em oposição à AC, cujo interesse, por mais ou menos três décadas,

estava relacionado quase exclusivamente ao nível verbal, como se reflete no sistema de Jefferson (2004). O termo "multimodal" pode enganar uma vez que sugere uma adição de modos já que a ideia de "múltiplas modalidades" carrega a semântica de canais separados que precisam ser reunidos embora se trate de uma relação intermodal que contribui para a abundância e a densidade da construção de significado. A rigor e epistemologicamente falando, o conceito da multimodalidade contrasta com uma compreensão mais holística, gestáltica e sinestésica de nossa experiência (Lyons, 2016; Deppermann, 2013).

Com este "renascimento" do corpo nas áreas da linguística, comunicação e interação, o processo de transcrição tem ganhado destaque como uma questão de confiabilidade. Uma vez que não é possível nem prático transcrever todos os aspectos de uma interação multimodal, os transcritores sempre se deparam com a escolha do que transcrever e do que deixar de fora, a que modos prestar atenção e qual nível de detalhes incluir (Lyons, 2016, p. 274). Além disso, os dados primários passam por uma redução inicial do original durante a gravação em vídeo, o que se torna mais relevante ainda em caso de gravações de interações virtuais, pois geralmente apenas o excerto virtual é gravado ao passo que o excerto do ambiente real, que inclui as outras partes corporais dos interlocutores, também não aparece na transcrição nem no vídeo. Sendo assim, o postulado de que é a posição da câmera que determina quais aspectos da realidade dada são selecionados para gravação de vídeo (Cruz *et al*., 2019) ganha uma importância enorme.

O surgimento da perspectiva multimodal trouxe mudanças significativas para as convenções de transcrição também. A transcrição passou de simplesmente converter a fala para a forma escrita para se tornar um esforço multifacetado e multicamadas, com uma preferência por formatos tabulares de transcrição em detrimento de sequências que podem parecer excessivamente centradas no conteúdo verbal para uma abordagem holística da comunicação multimodal. As transcrições tabulares geralmente segregam modos específicos, como olhar, gestos, vocalização e interações com artefatos e objetos, em colunas. Elas procuram manter informações sobre como esses modos interagem entre si, permitindo leituras verticais e horizontais para capturar a ocorrência simultânea de modos em um ponto específico no tempo, bem como o *layout* síncrono da linha do tempo (Cowan, 2014; Flewitt *et al*., 2017).

A questão de como lidar com dados ricos (*rich data*) tornou-se cada vez mais urgente, à medida que mais e mais conjuntos de dados estão disponíveis por meio de fontes *online* ou projetos de compilação multimodal. No decorrer dos últimos anos surgiram assim tentativas para elaborar novos caminhos para transcrever interações multimodais com base em convenções que representam diferentes níveis de recursos (Heath *et al*., 2010; Mondada, 2019). No entanto, até este ponto, nenhum formato atingiu o nível desejado de completude. Permanece desafiador fornecer uma representação adequada da intrincada simultaneidade de diferentes modos e representar de forma uniforme recursos visuais e corporais. Além disso, a complexidade excessiva de alguns sistemas de transcrição gera críticas de leitores e pesquisadores, que apontam a falta de

clareza e legibilidade. Bezemer e Jewitt (2010) apontam o dilema que os pesquisadores enfrentam: muita atenção a muitos modos diferentes pode prejudicar a compreensão dos significados de um modo específico; muita atenção a um único modo corre o risco de "fixar as coisas" em apenas uma das muitas maneiras pelas quais as pessoas atribuem significado. Além disso, não há uma maneira padrão de transcrever dados ricos. Sempre terá e deve ter uma adaptabilidade da transcrição aos dados específicos, à pergunta de pesquisa e ao objetivo de uso. Convém destacar que um critério importante é manter um alinhamento claro e uma representação de simultaneidade e orquestração de diferentes recursos interativos, bem como uma certa expansibilidade (princípio da cebola) da transcrição em relação a diferentes graus de detalhamento.

Como Schröder (no prelo) aponta, um caminho de saída poderia ser um procedimento em dois passos: elaborar uma transcrição sequencial clássica (GAT 2) como primeiro passo de uma sequência maior e elaborar um *zoom in* como segundo passo por meio de uma transcrição tabular para uma anotação de gestos. Nos EdG, por muitas décadas, houve um foco primário nos gestos manuais junto a certa exclusão de uma visão holística, não apenas de outros recursos como pistas prosódicas, mas particularmente também de outros tipos de gestos tais como gestos faciais. Ora, em interações virtuais, a maior parte do corpo da pessoa é "cortada" e, como abordamos em 2.1, concomitantemente, a distância entre os interlocutores é reduzida. Sendo assim, os gestos faciais ganham destaque, e o transcritor percebe que há um repertório bastante fragmentado disponível para lidar com esse novo universo de transcrição. Bavelas e Covil, que tratam de gestos faciais há mais de duas décadas, ainda constam em 2006 de que o estudo de gestos faciais apresenta "novas direções de pesquisa nesta área relativamente negligenciada"[6] (Bavelas; Chovil, 2006, p. 98) e destacam repetidamente que "os gestos faciais na conversa não são expressões emocionais"[7] (Bavelas; Chovil, 2018, p. 98), como muitas pesquisas sugerem. É por isso que direcionaremos nosso olhar na próxima seção justamente a este fenômeno específico para ilustrar (i) a relevância (particular) dos gestos faciais para a transcrição de interação virtual; (ii) sua conceitualização epistemológica e (iii) como elaborar, com base nisso, as categorias para a transcrição de *talking heads*.

Categorias e funções dos gestos faciais

A análise dos gestos faciais é fundamental para entender linguagem humana e pode atuar na percepção e na produção da comunicação ao produzir pistas de ênfase, (des)aprovação, (des)engajamento, negação, confirmação, reparo lexical etc. entre os participantes durante uma interação. No contexto de encontros *online* e síncronos, onde os participantes podem ver a si mesmos na tela, mas são incapazes de identificar com precisão a quem ou o quê estão vendo em seus dispositivos, gestos faciais como sorrisos,

[6] No original: "new directions of research in this relatively neglected area".

[7] No original: "conversational facial gestures are not emotional expresssions".

endereçamento de olhar e acenos de cabeça tendem a modificar consideravelmente a sistemática de adjacência e alocação de turnos interacionais. Além disso, a dinâmica gestual dessas categorias possibilita aos participantes comunicar disponibilidade para iniciar uma interação, sinalizar (des)alinhamento estrutural, manifestar afiliação social e criar sequências de (auto)reparo (Hjulstad, 2016; Fidelis de Paula, 2023). Como a falta de recursos visuais em interações virtuais pode apresentar desafios à análise, compreensão e interpretação precisa dos elementos conversacionais, a atenção aos gestos faciais torna-se ainda mais premente. Desse modo, esta seção tem como objetivo apresentar o panorama dos gestos faciais na transcrição da interação virtual, juntamente com sua conceitualização epistemológica e criação de categorias destinadas à transcrição de "cabeças falantes" (*talking heads*), com base nas categorias e funções referenciais e pragmáticas propostas por Bavelas (2022).

Kendon (2004) define "gestos" como ações que compõem um sistema integrado único, sendo produzidos no contexto da fala ou equivalentes a um enunciado completo. Destarte, eles contribuem para a compreensão da multimodalidade na expressão co-verbal. Em consonância com esse entendimento, os gestos faciais correspondem às ações que englobam movimentos ou configurações da cabeça e/ou dos músculos e membros da face, abrangendo áreas como sobrancelhas, olhos e boca, frequentemente caracterizados pela coordenação anatômica e funcional desses movimentos.

De maneira análoga aos gestos manuais, os gestos faciais compartilham diversas propriedades, sincronizam-se de maneira harmoniosa com a(s) palavra(s) que complementa(m) e suplementa(m) e destacam a interconexão entre componentes físicos e expressivos na dinâmica da interação. Apresentam, portanto, uma notável capacidade comunicativa, ao fornecerem informações semânticas, sintáticas e pragmáticas. Além disso, os gestos faciais, em conjunto com os demais gestos, estão fortemente ligados ao contexto de sua produção e ao espaço interacional (Mondada, 2013), isto é, aos arranjos corporais dos participantes situados e coordenados de maneira recíproca no espaço. Isso implica que tais gestos não conferem significados de forma isolada e, tampouco, correspondem a "expressões" emocionais dos indivíduos (Bavelas; Chovil, 2018; Bavelas, 2022) nas línguas de modalidade oro-auditiva. Este ponto é relevante uma vez que os campos da psicologia dominavam por muitas décadas os conceitos científicos e populares sobre "expressões faciais". A abordagem mais famosa das expressões faciais foi desenvolvida por Ekman e Friesen (1978). Ao pressupor que o rosto é o principal local de exibição de afetos, os autores propõem que existe um conjunto fixo de configurações musculares que correspondem a emoções inatas e universalmente reconhecidas. Isso é refletido em seu Sistema de Codificação de Ações Faciais,[8] que permite ao pesquisador elaborar uma saída de expressão facial com base na fusão de nove métricas emocionais (incluindo "raiva", "desprezo", "nojo", "medo", "sentimentalismo", "alegria", "tristeza", "surpresa" e "confusão") com vinte métricas de expressão facial baseadas na medição da ativação muscular facial, refletindo o

[8] https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics

envolvimento emocional do sujeito. É para destacar sua importância interacional e seu significado variável *in situ* que Bavelas e Chovil preferem o termo "gesto facial".

Por outro lado, os gestos faciais apresentam uma diferença significativa em termos de tempo e execução e, mais especificamente, não precisam ser alocados em uma posição exata dentro da unidade gestual ou mesmo ter sua realização subdivida em fases, a exemplo do que ocorre com os gestos manuais (Bressem; Ladewig; Müller, 2013), porque sua realização tende a ser extremamente breve. Para Bavelas e Chovil (2018), a peculiaridade desses gestos reside exatamente nesta capacidade de eles ocorrerem em frações de segundos e na possibilidade de estarem fortemente sincronizados com a fala. No entanto, isso não significa que os gestos faciais estejam necessariamente vinculados a um ponto específico da produção oral. Em um estudo sobre a contribuição dos gestos faciais para a percepção das funções pragmáticas – entoação multimodal e foco – em português brasileiro, Carnaval *et al*. (2023, p. 25) evidenciam que determinadas ocorrências de gestos faciais podem persistir ao longo de toda a produção da unidade conversacional, sem necessariamente estarem sincronizadas com um momento específico da fala.

Os trabalhos conduzidos por Chovil (1997, 2005), dedicados à análise do significado e à função de gestos faciais em determinadas ocorrências conversacionais, estabeleceram um alicerce para as categorias e funções delineadas por Bavelas (2022). Adotando uma abordagem indutiva, Chovil identificou quatro categorias gerais, cada uma delas abarcando até 12 funções específicas relacionadas aos gestos faciais. Posteriormente, a proposta de Bavelas (2022) traz uma classificação mais abrangente, contribuindo para a compreensão mais aprofundada dos gestos faciais, ao classificá-los em (i) reações pessoais (*personal reactions*), (ii) representações (*portrayal*), (iii) mímicas motoras (*motor mimicry*), e (iv) sorrisos, acenos de cabeça e orientações do olhar (*gaze*). Os gestos executados em cada uma dessas categorias desempenham funções específicas, enriquecendo o entendimento das complexidades associadas à linguagem gesto-facial em suas diversas manifestações nas interações conversacionais. Com base em uma compreensão situacional e interacional do "gesto facial", em oposição à "expressão facial", as funções não apenas se sobrepõem, mas também variam dependendo do contexto, que está sempre em fluxo contínuo. A primeira categoria proposta por Bavelas (2022), denominada "reações pessoais", é frequentemente relacionada à função referencial e representa uma das ocorrências mais comuns de gestos faciais. Ações como erguer as sobrancelhas, arregalar os olhos e deixar o queixo cair refletem a reação dos falantes à conversa em curso e desempenham diferentes papéis na interação, como enfatizar, dar retorno ou confirmação (*backchannel*), expressar pensamentos/lembranças, destacar um elemento interacional, oferecer/sugerir, ou ainda franzir o rosto (*facial shrug*) em um movimento equivalente a encolher/dar de ombros. Vale ressaltar que, embora os ouvintes também possam produzir gestos faciais de reações pessoais durante a interação, isso ocorre com menor frequência em comparação com os falantes.

As representações (*portrayal*) são realizadas para ilustrar uma cena, expressão ou reação previamente ocorrida, seja do falante ou de alguém que ele esteja descrevendo

durante a conversa. Tais gestos envolvem as variações nas expressões e outras categorias de gestos faciais, como o aceno da cabeça e a orientação do olhar. Esses gestos mantêm uma estreita relação com as funções referenciais da linguagem e podem indicar mudanças de frase, negação, foco social, conclusão do turno de fala, concessão ("mas"), sarcasmo ou humor.

Já os gestos faciais de mímica motora (*motor mimicry*) são executados majoritariamente por ouvintes, quando estes respondem à fala em curso como se estivessem se colocando no lugar do falante em um dado momento da narrativa (Bavelas, 2007). Esses gestos muitas vezes servem como uma reação empática, demonstrando que o ouvinte se conecta emocionalmente à experiência narrada, dando mostras de surpresa, espanto ou mesmo um leve susto diante do relato. Nesta classificação, ações como levantar as sobrancelhas, franzir o nariz ou arregalar levemente os olhos são reconhecidas por proporcionarem uma resposta adequada à expressão do falante e, portanto, podem desempenhar funções interacionais como conclusão de um tópico, encerramento conversacional, demonstração de compreensão ou concordância, explicação, mudança de assunto ou mesmo um sinal de pausa na conversa.

O sorriso, diferentemente das demais categorias, recebeu maior atenção nos EdG (cf. Goodwing; Goodwing, 2000; Kendon, 2004; Kaukomaa *et al*., 2013; Bavelas; Chovil, 2018), os quais identificaram uma variedade de funções em interações face a face. Estas incluem a capacidade de regular diferentes estágios da interação (Kendon, 2004), atuar em sinal de resposta e/ou confirmação intraturnos de modo semelhante ao aceno de cabeça (Bavelas, 2022), antecipar ou dar pistas sobre um conteúdo da conversa, destacar algum elemento na fala em curso, expressar humor, ironia ou autodepreciação (Bavelas; Chovil, 2018).

Os acenos de cabeça são comumente utilizados para expressar concordância ou aprovação, ou ainda, negação ou discordância. Além disso, os falantes podem utilizar o ângulo da cabeça para representar relações, enfatizar pontos específicos na conversa, evidenciar o papel e a perspectiva do falante, ou destacar uma direção específica. Os gestos produzidos com acenos de cabeça frequentemente se combinam com a orientação do olhar para enfatizar elementos específicos na comunicação e permitem aos falantes reforçar a clareza e o impacto de suas sugestões não oralizadas.

A orientação do olhar, por sua vez, atua tanto na regulação da tomada dos turnos de fala (Kendon, 1967; Fidelis de Paula, 2023), quanto no direcionamento da atenção e destaque aos gestos manuais, assume uma função dêitica (Hjulstad, 2016). A troca de olhares, mais especificamente, pode atuar na coordenação temporal dos turnos oralizados, seja ao dar mostras ou solicitar confirmação (*backchannel*).

Em resumo, a análise dos EdG, cujo progresso está intimamente associado aos propósitos comunicativos dos interlocutores, revela que a significância dos gestos faciais não está apenas em contextos presenciais, mas também em interações virtuais, influencia diretamente a exatidão da transcrição e a compreensão do processo comunicativo, particularmente no mundo global e em encontros interculturais, como abordado inicialmente. A seguir, serão apresentadas duas propostas de sistemas de transcrição

desenvolvidas recentemente: (a) GAT 2 em associação ao ISWA, de Dix (2022) – e (b) ViMELF, de Brunner e Diemer (2021), para ILF em contextos de videoconferência. Tal abordagem metodológica visa uma compreensão mais abrangente e estruturada dos gestos faciais e sua importância nas interações conversacionais virtuais, e buscam, deste modo, estabelecer ferramentas concretas e uma fundamentação consolidada metodologicamente para pesquisas futuras.

Fundamentação metodológica: duas propostas de sistemas de transcrição ilustradas

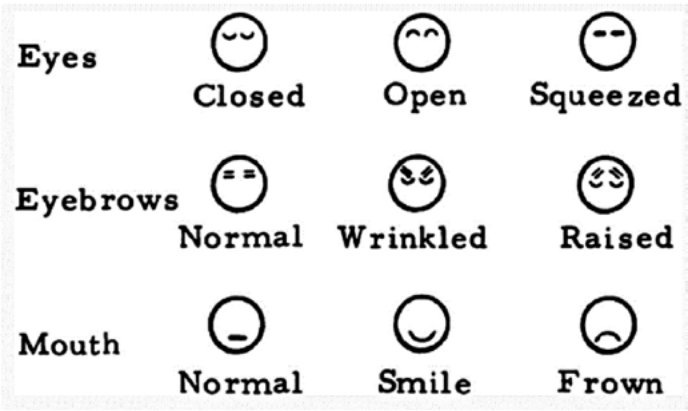
GAT 2 + ISWA (Dix 2022)

Com base no GAT 2, Carolin Dix (2022) recentemente desenvolveu uma proposta para combinar as convenções de transcrição para fala-em-interação com o inventário do International SignWriting Alphabet (ISWA) de Sutton (2010). A intenção da pesquisadora é oferecer uma ferramenta capaz de representar e integrar diferentes níveis de recursos corporais, uma vez que, como discutido anteriormente, sua crítica também se volta para o fato de que a maioria das transcrições multimodais se concentra em um único recurso, como gestos manuais ou olhares. Por outro lado, Dix também destaca que essa integração requer adaptabilidade do transcrito à correspondente base de dados, ao objetivo da pesquisa e ao seu uso. Em termos gerais, isso significa que há a possibilidade de uma integração detalhada e precisa de todos os fenômenos, bem como sua orquestração e simultaneidade, mas existe, assim como no GAT 2, o princípio da "casca de cebola", no qual as transcrições são realizadas em três níveis distintos de granularidade (Schröder; Nascimento; Silva, 2019). Quanto à notação dos aspectos corporais no caso da ISWA, ela pode ser expandida ainda mais, de acordo com o foco da pesquisa (Dix, 2022).

O aspecto mais relevante dos pictogramas do ISWA é seu *status* universal. Desenvolvido nos anos setenta por Valerie Sutton (2010), ele é composto por ícones simbólicos para a notação da língua de sinais. As notações incluem a perspectiva do produtor e compreendem cinco categorias: (a) mãos, (b) movimentos, (c) dinâmicas, (d) cabeça e rostos, e (e) corpo (https://www.signbank.org/iswa/). Dix (2022) ilustra como é criada uma transcrição de partitura, a partir da inserção de linhas visuais abaixo da linha verbal (GAT 2) em correspondência com os diferentes níveis de recursos (orientação corporal, posicionamento da cabeça, posição da mão, movimento dos ombros, gestos faciais etc.). Sem entrar em detalhes aqui, para resumir genericamente, a dificuldade da legibilidade dos símbolos e sua integrabilidade em um sistema de transcrição da fala-em-interação já reside no fato de que, no total, há 261 símbolos e esses símbolos nem sempre são facilmente transparentes para uma pessoa leiga no assunto.

Ao estudarmos os símbolos do ISWA e a maneira como Dix os integra em sua proposta de transcrição multimodal, chamou nossa atenção o amplo repertório de símbolos para recursos faciais e sua consonância com as formas e funções dos gestos faciais estabelecidos por Bavelas (cf. Categorias e funções dos gestos faciais). Essa categoria implica símbolos para as seguintes subcategorias: (a) movimento de cabeça e direção do rosto; (b) formas das sobrancelhas, olhos e direção do olhar; (c) bochechas, nariz, orelhas e respiração; (d) forma da boca e dos lábios; e (e) língua, dentes, queixo e pescoço. Alguns símbolos de fácil compreensão em termos icônicos podem ser vistos na Figura 1.

Figura 1 – Símbolos para (a) olhos (fechados, abertos, espremidos), (b) sobrancelhas (normais, enrugadas, levantadas), (c) boca (normal, sorriso, carranca)



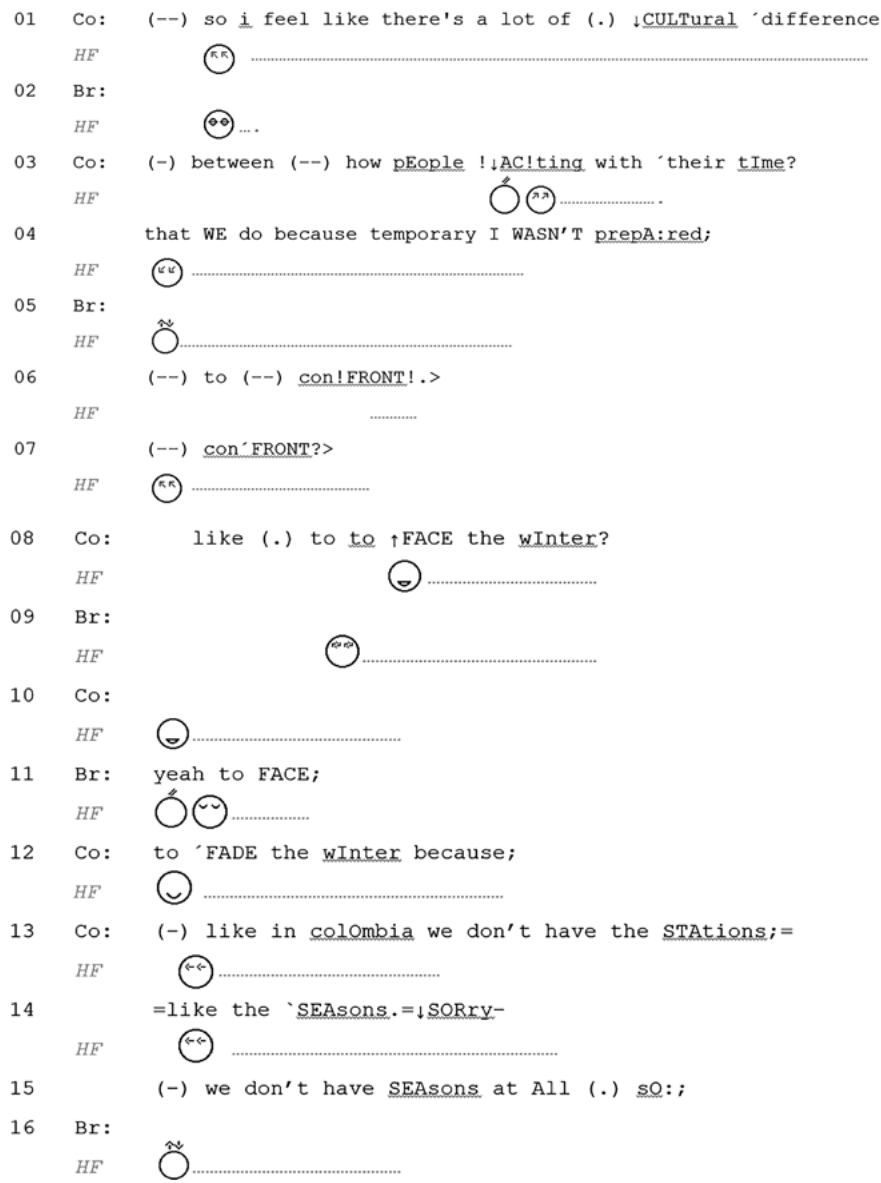
Fonte: Sutton (1982, p. 82 *apud* Dix, 2022, p. 114)

A seguir, apresentaremos a transcrição da sequência introduzida de acordo com alguns símbolos do inventário da ISWA. Ainda estamos em uma fase em que nosso aparato metodológico não está concluído, mas após algumas tentativas, decidimos que, para nossa finalidade e para manter a legibilidade máxima e o princípio da relevância, manteremos, para uma primeira discussão, apenas um nível de transcrição para todos os recursos corporais, em vez de incluir uma transcrição de partitura.[9] A primeira linha mostra a transcrição de acordo com as convenções de GAT 2 (Schröder *et al.*, 2016) junto às siglas da origem das pessoas, de acordo com as normas do ICMI uma vez que se trata de uma conversa em ILF *(Co=Colombian; Br=Brazilian)* ao passo que a segunda linha mostra as atividades dos gestos faciais junto ao tempo de execução (*HF=Head; Face*).

[9] Para ver o formato de uma transcrição com vários níveis junto à inserção de (potencialmente) todas as categorias de símbolos e particularmente para um entendimento da transcrição em GAT 2 que não será discutida aqui, ver anexo.

Figura 2 – Transcrição da Sequência 1 de acordo com GAT 2 + ISWA (DIX, 2022)

Sequência 1: *To face the winter* **2019SkyBrCo01 ((07:13-07:36))**[10]



Fonte: Elaboração própria

[10] O vídeo pode ser acessado no seguinte *link*: https://youtu.be/t-wjyou1rtE

A seguir, iremos brevemente exemplificar a leitura e possível entendimento de alguns ícones para ilustrar como nossa adaptação do sistema de Dix (2022) para a transcrição multimodal de fala-em-interação mediada por vídeo procede:

Nas linhas 01, 03, 07, 13 e 14, observa-se que o colombiano (Co), nesses momentos específicos, junto à fala, ao invés de direcionar seu olhar para a tela, desvia seu olhar para cima e para a direita, ou esquerda, ou apenas para a esquerda. Isso indica prototipicamente um acesso à memória ao invés de um endereçamento ao interlocutor, o que corresponde ao nível verbal no qual haja, simultaneamente, inserções metaenunciativas. Por exemplo, nas linhas 06 a 07, ele não sabe se escolheu o termo correto, indicado pela inserção de marcadores de hesitação e pausas. Ao olhar para cima e à esquerda, ele exibe um "rosto pensando" (*thinking face*, Bavelas; Chovil, 2018), como se buscasse pela palavra em outro espaço cognitivo. Em linha 13, ao olhar à esquerda, ele utiliza o termo *stations*, provavelmente como uma transferência do espanhol *estaciones* para *seasons*, mas imediatamente apresenta um autorreparo iniciado por si mesmo: `"=like the "SEAsons."↓SORry-` (linha 14). Aqui, a unidade de construção de turno é momentaneamente suspensa (Hayashi; Raymond; Sidnell, 2013, p. 13). Os autorreparos são técnicas altamente relevantes em comunicação intercultural (Liu; Kinginger, 2021; Cogo; House, 2018).

Já os símbolos corporais que acompanham o comportamento do interlocutor brasileiro (Br) também indicam, à primeira vista, um alto engajamento na co-construção do significado e na afiliação. Na linha 02, o piscar dos olhos sinaliza concentração e disposição, enquanto nas linhas 05 e 16 ele acena com a cabeça. O mais interessante é que os símbolos nas linhas 09 e 11 de certa forma permitem um acesso ao plano cognitivo e sua co-construção ativa na busca do lexema correto. Na linha 09, o brasileiro para o seu acompanhamento rítmico da conversa para também acessar sua própria memória quando os olhos se movem de cima para a esquerda e voltam para cima, visualizado pelo ícone. Posteriormente, na linha 11, ele acena com a cabeça para a esquerda e fecha os olhos ao dizer `yeah to FACE;`. É justamente neste contexto que os dois sorrisos com a boca levemente aberta são exibidos pelo colombiano, talvez em sinal de que ele seja grato por receber ou ter recebido apoio de seu interlocutor (linhas 08, 10 e 12).

Sendo assim, os símbolos icônicos do ISWA, integrados aqui da categoria "cabeça e rosto", de forma ainda mais simplificada do que proposto por Dix (2022), ajudam o pesquisador e o leitor a identificar, já em um primeiro momento, até que ponto uma sequência específica de uma interação se torna relevante e significativa para a análise, ao fornecer pistas para a direção que poderia ser explorada mais.

ViMELF (Brunner; Diemer, 2021)

O desenvolvimento de um sistema de notação para a fala-em-interação no *corpus* ViMELF (*Video-mediated English as a Lingua Franca Conversations,* 2018) ressaltou

a relevância de transcrever os elementos não verbais (ENV)[11] salientes, os quais incluem gestos, expressões faciais, olhares, postura física, assim como mudanças de câmera e eventos de fundo, de maneira breve e estruturada, conforme enfatizado por Brunner e Diemer (2021). O objetivo principal dos autores é fornecer um modelo de transcrição robusto e conciso, projetado para integrar eficientemente dados multimodais no *corpus* do ViMELF. Este *corpus* compreende conversas informais gravadas entre participantes de várias nacionalidades e desconhecidos entre si, engajados em interações *online* e síncronas mediadas pelo Inglês como Língua Franca. Além disso, espera-se que esse modelo seja aplicável em diversas áreas, sem a necessidade de um conhecimento prévio em pesquisa de gestos. Embora reconheçam a existência de certa subjetividade na definição dos critérios de saliência e concisão, a abordagem proposta oferece flexibilidade para analisar uma ampla gama de dados multimodais e permite investigações tanto quantitativas quanto qualitativas das interações conversacionais.

Por outro lado, Brunner e Diemer (2021) apontam que, apesar do considerável histórico e da influência dos Estudos de Gestos (EdG) no avanço das pesquisas com dados multimodais, grande parte dos trabalhos realizados até então concentrou-se na descrição de gestos e, por conseguinte, negligenciou a transcrição efetiva desses dados. Este dilema inicial na descrição de gestos é atribuído principalmente à complexidade das notações, que frequentemente se baseiam em elementos visuais para a classificação dos gestos, o que dificulta as análises quantitativas. Além disso, a transcrição tabular em múltiplos níveis e ilustrada (Goodwin, 2007; Mondada, 2018; Due; Lange, 2021) representa um desafio considerável, tanto pela sua natureza detalhada e descritiva, quanto pela complexidade envolvida na sua interpretação. Ademais, os autores destacam que as abordagens centradas na dinâmica interacional implicam um grau significativo de interpretação por parte dos transcritores.

Em contraste com a transcrição da AC, cuja aplicabilidade em pesquisas quantitativas é limitada, o sistema proposto por Brunner e Diemer (2021) integra tanto a anotação gestual quanto demais elementos, como pausas, marcadores de hesitação, respiração etc. no texto geral da transcrição, em vez de criar uma camada separada para os elementos não verbais.

A transcrição dos ENV salientes segue um padrão específico, onde as ações são colocadas entre chaves curvas {} com o(s) verbo(s) conjugado(s) na terceira pessoa do singular, sempre no presente do indicativo, como `{sorri}`, não `{sorrindo}`. Para ENV consecutivos, a notação é sequencial, `{levanta a cabeça}, {revira os olhos}`, enquanto ocorrências simultâneas são indicadas pelo símbolo "&" `{levanta sobrancelhas & sorri}`. Tais notações não se relacionam a um conteúdo entonacional específico nem são marcadas em relação à velocidade, intensidade ou duração da execução, o que apresenta um ponto fraco dessa proposta. Além disso, é possível ampliar o escopo da transcrição para incluir (i) objeto direto, para verbos que não trazem implicitamente o objeto (levantar *a mão*, erguer *o braço*),

[11] Do original, o termo *Nonverbal Elements* (NVE). Tradução livre.

diferentemente de {assentir} que inclui o domínio "cabeça"; (ii) advérbio (três vezes, repetidamente); (iii) preposições (*para* cima, *à* esquerda, *com* o dedo); (iv) o sufixo "-ndo" para indicar modificações adicionais como movimentos feitos pelos participantes e, por fim, se houver necessidade de incluir mais detalhes, (v) as ações podem ser separadas por ponto e vírgula {levanta as mãos; palmas para fora}, {levanta a mão; palma para cima}, sempre com o princípio de concisão em mente, para que as transcrições permaneçam o mais legíveis possível.

Na sequência, vamos exemplificar a notação e leitura de alguns elementos salientes na interação para ilustrar nossa adaptação do sistema de Brunner e Diemer (2021) para a transcrição multimodal de fala-em-interação mediada por vídeo:

Sequência 1: *To face the winter* **2019SkyBrCo01 ((07:13-07:36))**

```
01  Co:   (--)[so i feel                ] like there"s a lot of
          (.) ↓CULTural ´difference, { olha para cima à direita }
02  Br:       [{pisca, acena com a cabeça}]
03  Co:   (-) between (--) how pEople !↓AC!ting with ´their tIme?
          {executa gesto manual & olha para a tela}
          { inclina a cabeça & olha para cima à direita }
04        that WE [do because temporary I WASN"T prepA:red;]
          { olha para baixo à esquerda } {olha para a tela}
05  Br:             [{acena com a cabeça}                    ]
06  Co:   (--) to (--) con!FRONT!.
07        (--) con´FRONT? {olha para cima à esquerda}
08  Co:   (--) like (.) to [to ↑FACE] the wInter? {olha para a tela
          & sorri}
09  Br:                    [to;     ]
10        {revira os olhos para cima e à esquerda e de novo
          para trás}
11  Co:   {sorri}
12  Br:   yeah to FACE; {inclina a cabeça para a direita & fecha
          os olhos}
13  Co:   to ´FACE the wInter because; {sorri & olha à esquerda}
14  Co:   (-) like in colOmbia we don"t have the STAtions;=
          {sorri & olha para a esquerda}
15        =like the `SEAsons.=↓SORry- {acena & olha à esquerda}
```

```
16          (-) we don”t have SEAsons [at All (.)              ] sO:;
            {olha para a tela}
17  Br:     [{acena com a cabeça}  ]
```

Nas linhas 01 e 02, podemos observar uma sobreposição dos turnos dos participantes, onde a unidade entonacional do colombiano começa `so i feel like there”s a lot of (.) ↓CULTural´difference` (linha 01), enquanto o brasileiro realiza gestos faciais sutis, como piscadelas e leves acenos de cabeça (linha 02). Esses gestos do brasileiro ocorrem simultaneamente ao seu olhar endereçado à tela e sugerem um elevado grau de envolvimento na construção do alinhamento interacional e na promoção da afiliação social (Stivers, 2008; Fidelis de Paula, 2023). Essa sobreposição de turnos multimodais é uma adaptação que desenvolvemos e que se diferencia do sistema proposto por Brunner e Diemer (2021), o qual preconiza apenas a inclusão do ENV, sem especificar a relação temporal ou contextual entre os gestos e o conteúdo verbal. Entretanto, entendemos que a associação dos gestos com elementos específicos da conversação pode indicar diferentes estratégias de construção de significado na interação. Além disso, a análise dessas ocorrências simultâneas pode proporcionar *insights* sobre a relação entre gestos faciais e aspectos narrativos e sequenciais, tanto no contexto IMV quanto no uso de ELF pelos participantes.

Nas linhas 06 e 07, ao olhar à esquerda, o colombiano utiliza o verbo *to confront*, provavelmente como uma transferência do espanhol *confrontar* para *to face*, mas logo na sequência faz um autorreparo iniciado por si: `(--) like (.) to [to ↑FACE] the wInter?` (linha 08), ao mesmo tempo em que endereça seu olhar diretamente para frente, possivelmente voltado para a câmera e sorri. Isso sugere que, depois de acessar o espaço cognitivo para realizar essa busca lexical, ele retorna ao ambiente interacional, olhando para a tela e se dirigindo ao brasileiro, ao mesmo tempo em que exibe um sorriso. Esse comportamento indica que o colombiano alterna entre o espaço cognitivo, onde realiza operações mentais, e o espaço interacional da tela, onde faz ajustes conversacionais e sinaliza engajamento com o brasileiro. O sorriso ao olhar para frente e possivelmente para a câmera parece marcar essa transição de volta à interação virtual e síncrona (Fidelis de Paula, 2023).

O sistema proposto por Brunner e Diemer (2021) permite a transcrição concisa de dados não verbais (ENV), simplificando a leitura das transcrições e fornecendo pistas imediatas sobre a relevância desses elementos interacionais para a análise. A taxonomia resultante desse modelo é baseada nos princípios de saliência e concisão, constituindo assim um sistema de anotação organizado, descritivo e abrangente. Entendemos que essa abordagem representa uma solução equilibrada, uma vez que permite que os pesquisadores estruturem dados multimodais complexos e contribuam para o desenvolvimento de *corpora* de dados ricos e uma ampla gama de aplicações.

Discussão

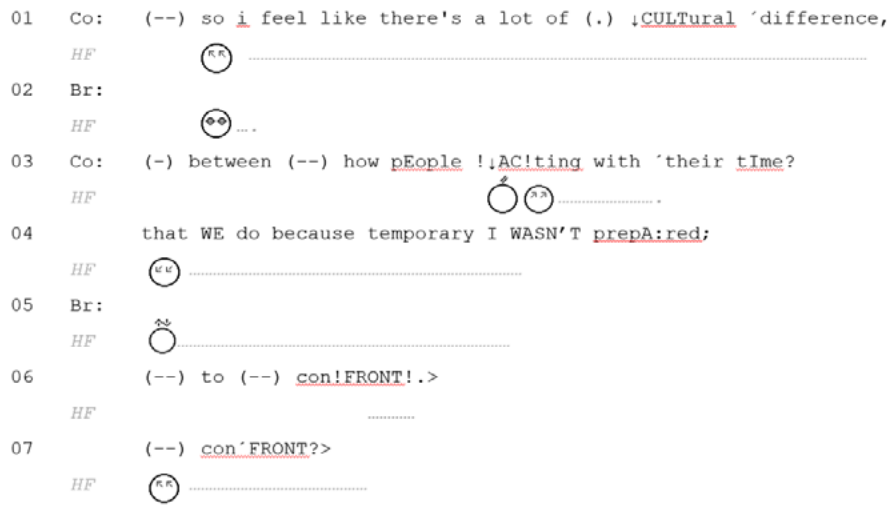
Uma primeira questão que emerge ao tratar dos desafios em relação à transcrição em interações mediadas por vídeo (IMV) diz respeito ao entendimento e à delimitação da saliência dos elementos gestual-corporais que devem ser transcritos. Essa tarefa nos leva a refletir particularmente sobre a natureza dinâmica e multifacetada dos gestos faciais em interações IMV em que alguns gestos, como a movimentação e o endereçamento do olhar, podem ser dificilmente estabelecidos com precisão. Nem sempre é possível determinar quando um movimento do olhar se torna um gesto relevante para a interação, dada a brevidade extrema de sua execução. Além disso, essa rapidez muitas vezes dificulta relacionar os gestos faciais a trechos específicos da fala. A seguir, discutimos como os modelos de transcrição propostos por Brunner e Diemer (2021) e Dix (2022) foram aplicados para analisar o direcionamento e o endereçamento do olhar, e destacamos os aspectos analíticos relacionados à notação de gestos faciais.

Sequência 1: *To face the winter* **2019SkyBrCo01 ((07:13-07:36))**

Modelo de transcrição Brunner e Diemer (2021)

```
01  Co:  (--) [so i feel                ] like there"s a lot of
         (.) ↓CULTural ´difference, {olha para cima à direita}
02  Br:      [{pisca, acena com a cabeça}]
03  Co:  (-) between (--) how pEople !↓AC!ting with ´their tIme?
         {executa gesto manual & olha para a tela}
         {inclina a cabeça & olha para cima à direita}
04       that WE [do because temporary I WASN"T prepA:red;]
         {olha para baixo à esquerda} {olha para a tela}
05  Br:         [{acena com a cabeça}                  ]
06  Co:  (--) to (--) con!FRONT!.
07       (--) con´FRONT? {olha para cima à esquerda}
```

Modelo de transcrição Dix (2022)



O sistema de Dix (2022) reduz as frases movimentais enquanto busca representar um modo preciso e lúcido para o tipo da ação. Vimos que o sistema de Dix (2022) traz uma vantagem com relação à introdução de uma universalidade dos ícones propostos, o que apresenta um aspecto de relevância crescente para a transcrição de VMI perante seu caráter cada vez mais multilíngue e intercultural em um mundo global, particularmente neste âmbito interacional. Ademais, como foi ilustrado a partir do nosso exemplo, o transcritor já percebe visualmente de modo imediato onde na interação ocorrem momentos intensos e multimodais de coconstrução de sentido, de compreensão recíproca e de possíveis incongruências. Sendo assim, a simplificação da transcrição mediante a adoção de símbolos icônicos oferece uma abordagem acessível e eficiente para a identificação de gestos faciais em uma versão reduzida e mais concisa, voltada para IMV, como proposta aqui. Essa simplificação não apenas acelera o processo de transcrição, como também auxilia o transcritor a desenvolver uma compreensão mais clara dos critérios de saliência dos gestos faciais. Os gestos são inseridos exatamente no ponto em que começam e sua duração é assinalada por meio de sinais de pontuação, como pontos (.), indicando o prolongamento da execução gestual em relação à fala. Porém, um ponto negativo no sistema da Dix não é apenas o esforço enorme que uma transcrição necessita, inclusive o treinamento para poder se tornar fluente na transcrição, mas também o fato de que o leitor de uma transcrição geralmente não está familiarizado com os símbolos.

Já a proposta de Brunner e Diemer (2021) busca oferecer uma transcrição padronizada que integra não apenas diretrizes semânticas nítidas, mas também considera o movimento integral em si de forma mais precisa do que os ícones. Sendo assim, tudo poderia ser documentado. Vimos que o comportamento do colombiano sugere uma tendência a desviar o olhar para cima e para a direita ou esquerda em vários momentos. Contudo, o ato de direcionar o olhar diretamente para frente (linhas 03 e 04), notado

exclusivamente no modelo de Brunner e Diemer (2021), reintroduz o colombiano no contexto da interação virtual e o coloca novamente no alinhamento social e para a organização da tomada de turnos de fala neste ambiente (Fidelis de Paula, 2023). Este *default case* não é previsto nos ícones mais estáticos da Dix. Da mesma forma, o sistema de Dix não permite anotar o olhar integral do colombiano na L03, ou seja, o objeto de interesse, o próprio gesto manual, está fora do foco do ícone por não mostrar a relação entre o apontamento do olhar e o gesto manual. Dado que o apontamento do olhar parece assumir, neste contexto, uma função mais autorreferencial do que interativa, a impossibilidade de registrar essa dinâmica pode restringir a compreensão do pesquisador sobre a construção de significados na interação mediada por vídeo. Assim, a análise detalhada e sensível ao contexto se mostra fundamental para compreender a dinâmica dos gestos faciais e do endereçamento do olhar e sua contribuição para a interação comunicativa.

A formalização dos elementos não verbais adicionalmente permite àqueles que buscam aplicar pesquisas quantitativas em seguida uma base sólida para a realização disso. Brunner e Diemer oferecem uma taxonomia rudimentar que se baseia em uma sintaxe simplificada para que possam ser criados *corpora* multimodais pesquisáveis. Esta possibilidade ainda não existe com relação a ícones, embora haja primeiros projetos para a integração de sistemas disso em programas de transcrição tais como EXMARaLDA (Schmidt; Wörnder, 2009). Essa abordagem mais detalhada e flexível pode ser particularmente útil quando se trata de compreender nuances importantes e complexidades das interações conversacionais. Além disso, a sintaxe concisa e padronizada desse sistema parece ser a melhor saída, até o momento, para retomar metadados e fazer buscas em *corpus* de dados ricos, sendo compatível com *softwares* que possuem sistemas de criação de banco de dados e ferramentas de buscas, como o Exakt (Schmidt; Wörner, 2014). Essa estrutura padronizada também se mostra favorável à integração com ferramentas de Inteligência Artificial (IA) que realizam transcrição automática (https://exmaralda.org/en/).

Entretanto, a observação do direcionamento e apontamento do olhar pelo colombiano sugere um processo autorreferencial, onde ele executa o gesto mais para monitorar e processar sua própria fala do que para direcioná-lo explicitamente ao interlocutor. Este aspecto aponta para a necessidade de um sistema de transcrição mais flexível, como proposto por Brunner e Diemer (2021), que permita a anotação detalhada dessa relação entre modalidades. Isso pode incluir mudanças de angulação e movimento de câmera; eventos de pano de fundo (sons, pessoas e/ou animais visíveis na interação, interrompendo ou interagindo com um dos falantes, etc.). Esse enfoque pode ser particularmente relevante para analisar interações onde o gesto facial serve tanto para expressar quanto para internalizar significados durante a comunicação mediada por vídeo.

Conclusão

Iniciadas nas últimas décadas, as pesquisas com dados multimodais ainda apresentam muitas limitações quando se trata de analisar os gestos faciais. Esse artigo se propôs a apresentar ferramentas metodológicas e de análise para transcrição de gestos faciais em IMV a partir dos modelos de transcrição propostos por Brunner e Diemer (2021) e Dix (2022). Escolhemos as duas propostas por dois motivos principais: (a) elas se inserem no debate atual sobre a complexidade e os problemas relacionados à transcrição multimodal no âmbito da Análise da Conversa, da Linguística Interacional e dos Estudos de Gestos; e (b) ambas tratam de questões relevantes para interações mediadas por vídeo (IMV), uma vez que uma abordagem oferece ferramentas icônicas universais para conversas em línguas francas, com um inventário detalhado para gestos faciais, enquanto a outra, ao apresentar uma proposta quase oposta, foi desenvolvida com base em IMV em ILF, buscando uma sintaxe concisa como base para pesquisas quantitativas. Observamos que, enquanto o sistema de Brunner e Diemer (2021) possibilita a notação mais detalhada e ampla dos gestos corporais, manuais e faciais, dos olhares e do contexto situacional da interação, o modelo de Dix (2022), feito a partir da inserção de símbolos icônicos adaptados do ISWA para cabeça e rosto, entrega uma transcrição concisa, evidenciando os gestos faciais mais salientes nas sequências interacionais. Nossa discussão aponta os desafios e as limitações em ambos os sistemas, uma vez que, em Dix (2022), é preciso um treinamento para criar familiaridade com os ícones e não há, até o momento, a possibilidade de inserir automaticamente os símbolos em *softwares* de transcrição, tais como EXMARaLDA, ELAN, CLAN etc. já o modelo de Brunner e Diemer (2021) não prevê a marcação temporal entre produção oral e gestual. Visando sanar essa lacuna em Brunner e Diemer (2021), propusemos adaptações para a notação de sobreposição de ações dos participantes, tanto para a fala, como gestos, aceno de cabeça, olhares e sorrisos, viabilizando a associação desses fenômenos a um momento específico da fala.

Entendemos que com a expansão dos *corpora* multimodais, incluindo aqueles que trazem uma variedade de contextos culturais e linguísticos, é essencial desenvolver sistemas híbridos, integrando a clareza visual dos ícones de Dix com a anotação detalhada de Brunner e Diemer. O desenvolvimento dessas ferramentas metodológicas e analíticas pode contribuir para o desenvolvimento de pesquisas futuras sobre a investigação da função autorreferencial dos gestos faciais em interações mediadas por vídeo, a relação entre o direcionamento e o apontamento do olhar com gestos manuais, bem como compreender como esses gestos são utilizados pelos interlocutores para monitorar e processar sua própria fala.

Agradecimentos e número CEP

Ulrike Schröder agradece à FAPEMIG, pelo fomento do projeto FAPEMIG Universal (2024-2026), a CAPES-DAAD, pelo fomento de PROBRAL (2023-2026)

e ao CNPq, pela Bolsa de Produtividade (2022-2025). Flavia Fidelis de Paula agradece à CAPES pelo apoio financeiro através do projeto CAPES/PROBAL, que possibilitou a realização da pesquisa de pós-doutorado na Universidade de Potsdam, na Alemanha, de novembro de 2023 a agosto de 2024, bem como à FAPEMIG, pela bolsa BDCTI, Nível I (2024-2025), no âmbito do projeto FAPEMIG Universal. O projeto "A coordenação multimodal da interação intercultural mediada por vídeo" foi aprovado pelo Comitê de Ética em Pesquisa pelo número do processo 55218521.1.1001.5149.

SCRÖDER, Ulrike; FIDELIS DE PAULA, Flavia. Video-mediated talk-in-interaction: challenges for the transcription of facial gestures. **Alfa**, São Paulo, v. 69, 2025.

- *ABSTRACT: The pandemic has restricted face-to-face interactions while accelerating the globalization of communication through video calls. This has facilitated the widespread conduct of activities such as collaborative work, scientific conferences, online teaching, medical consultations, religious rites, and family gatherings. These practices have not only promoted the expansion of local boundaries but have also connected the world on a global scale, enhancing intercultural encounters increasingly characterized by multilingual conversations. In this new configuration, "talking heads" assume central importance. This paper presents and compares two methodological models for transcribing facial gestures in video-mediated interactions, as proposed by Brunner and Diemer (2021) and Dix (2022). Using a multimodal conversation analysis approach, we apply these models to a video-mediated interaction (VMI). We will demonstrate the need to develop methodologies that capture the multimodal complexity of facial gestures in these interactions, highlighting how these systems can contribute to both quantitative and qualitative research methods integrate linguistic and gestural dimensions. Additionally, we will discuss the advantages and disadvantages of the two systems and propose enhancements.*

- *KEYWORDS: Multimodal transcription; Video-mediated interaction (VMI); Facial gestures; English as a lingua franca; Intercultural communication.*

Contribuição dos autores (conforme taxonomia CRediT)

Ulrike Schröder: Conceitualização, Curadoria de dados, Análise de dados, Recebimento de Financiamento, Pesquisa, Metodologia, Administração do projeto, Redação do manuscrito original, Redação - revisão e edição

Flavia Fidelis de Paula: Conceitualização, Análise de dados, Pesquisa, Metodologia, Redação do manuscrito original, Redação - revisão e edição

Declaração de Disponibilidade de Dados

Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

REFERÊNCIAS

BALAMAN, U.; DOEHLER, S. Navigating the complex social ecology of screen-based activity in video-mediated interaction. **Pragmatics**, v. 32, n. 1, p. 54-79, 2022.

BAVELAS, J. Face-to-face dialogue as a micro-social context. *In*: DUNCAN, S. D.; CASSELL, J.; LEVY, E. T. (org.). **Gesture and the dynamic dimension of language**. Amsterdam, The Netherlands: Benjamins, 2007. p. 127-146.

BAVELAS, J. **Face-to-face dialogue**: Theory, research, and applications. Oxford University Press, 2022.

BAVELAS, J.; CHOVIL, N. **Nonverbal and verbal communication**: Hand gestures and facial displays as part of language use in face-to-face dialogue, 2006.

BAVELAS, J.; CHOVIL, N. Some pragmatic functions of conversational facial gestures. **Gesture**, v. 17, n. 1, p. 98-127, 2018.

BAVELAS, J.; GERWING, J.; HEALING, S. Hand and facial gestures in conversational interaction. *In*: HOLTGRAVES, T. M. (org.). **The Oxford handbook of language and social psychology**. Oxford: Oxford University Press, 2014. p. 111-130.

BEZEMER, J.; JEWITT, C. Multimodal analysis: Key issues. *In*: LITOSSALIS, L. (org.). **Research methods in linguistics**. London: Continuum, 2010. p. 180-197.

BRESSEM, J.; LADEWIG, S.; MÜLLER, C. A linguistic annotation system for gesture. *In*: MÜLLER; C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (orgs.). **Body – language – communication**. An international handbook on multimodality in human interaction. Volume 1, Berlin, Boston: De Gruyter Mouton, 2013. p. 483-501.

BRUNNER, M.; DIEMER, S. Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription. **Research in Corpus Linguistics**, v. 9, n. 1, p. 63-88, 2021.

CARNAVAL, M.; MIRANDA, L. S.; MORAES, J. A.; RILLIARD, A. Funções dos gestos faciais na prosódia audiovisual. *In*: AVELAR, M.; PACHECO, V.; OLIVEIRA, M. (org.). **Linguística e Estudos de Gestos**: Interfaces. Campinas: Pontes Editores, 2023. p. 11-49.

CHOVIL, N. Facing others: A social communicative perspective on facial displays. *In*: RUSSELL, J. A.; FERNÁNDEZ-DOLS, J. M. (org.). **The psychology of facial expression**. Cambridge: Cambridge University Press, 1997. p. 321-333.

CHOVIL, N. Measuring conversational facial displays. *In*: MANUSOV, V. L. (org.). **Sourcebook of nonverbal measures***: Going beyond words.* Mahwah, NJ: Lawrence Erlbaum Associates, 2005. p. 173-188.

COGO, A.; HOUSE, J. The pragmatics of ELF. *In*: **The Routledge handbook of English as a lingua franca.** *In*: JENKINS, J.; BAKER, W.; DEWEY, M. (org.). London: Routledge, 2018. p. 210-223.

COUPER-KUHLEN, E.; SELTING, M. Towards an interactional perspective on prosody and a prosodic perspective on interaction. *In*: COUPER-KUHLEN, E.; SELTING, M. (org.). **Prosody in conversation**. Cambridge: Cambridge University Press, 1996. p. 11-56.

COWAN, K. Multimodal transcription of video: Examining interaction in Early Years classrooms. **Classroom Discourse**, v. 5, n. 1, p. 6-21, 2014.

CRUZ, F. M. D.; OSTERMANN, A. C.; ANDRADE, D. N. P.; FREZZA, M. O trabalho técnico-metodológico e analítico com dados interacionais audiovisuais: a disponibilidade de recursos multimodais nas interações. **DELTA**: Documentação de Estudos em Linguística Teórica e Aplicada, v. 35, 2019. Disponível em: https://doi.org/10.1590/1678-460X2019350404. Acesso em: 16 ago. 2024.

DEPPERMANN, A. Multimodal interaction from a conversation analytic perspective. **Journal of Pragmatics**, v. 46, n. 1, p. 1-7, 2013.

DIX, C. GAT2 trifft das International SignWriting Alphabet (ISWA). *In*: SCHWARZE, C.; GRAWUNDER, S. (org.). **Transkription und Annotation gesprochener Sprache und multimodaler Interaktion:** Konzepte, Probleme, Lösungen. Tübingen: Narr Francke Attempto, 2022. p. 103-131.

DUE, B. L.; LANGE, S. B. Body part highlighting: Exploring two types of embodied practices in two sub-types of showing sequences in video-mediated consultations. **Social Interaction. Video-Based Studies of Human Sociality**, v. 3, n. 3, 2021. Disponível em: https://doi.org/10.7146/si.v3i3.123836. Acesso em: 24 ago. 2024.

DUE, B.; LICOPPE, C. Video-mediated interaction (VMI): Introduction to a special issue on the multimodal accomplishment of VMI institutional activities. **Social interaction. Video-based studies of human sociality**, 3. Jg., Nr. 3, 2021. Disponível em: https://doi.org/10.7146/si.v3i3.123836. Acesso em: 24 ago. 2024.

EKMAN, P.; FRIESEN, W. V. **Facial Action Coding System**: A technique for the measurement of facial movement. Palo Alto: Consulting Psychologists Press, 1978.

FIDELIS DE PAULA, F. **A multimodalidade em aulas síncronas:** um estudo sobre a construção de afiliação e alinhamento no contexto de ensino remoto. 2023. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2023.

FLEWITT, R.; HAMPEL, R.; HAUCK, M.; LANCASTER, L. What are multimodal data and transcription? *In*: JEWITT, C. (org.). **The Routledge handbook of multimodal analysis**. New York: Routledge, 2017. p. 44-59.

FRIESEN, N. Telepresence and tele-absence: A phenomenology of the (in) visible alien online. **Phenomenology & Practice**, v. 8, n. 1, p. 17-31, 2014.

GOODWIN, C. Environmentally coupled gestures. *In*: DUNCEN, S.; CASSELL, J.; LEVY, E. (org.). **Gesture and the dynamic dimensions of language**. Amsterdam, The Netherlands: John Benjamins, 2007. p. 195-212.

GOODWIN, C. Why multimodality? Why co-operative action? **Social Interaction. Video-Based Studies of Human Sociality,** v. 1, n. 2, 2018. DOI: https://doi.org/10.7146/si.v1i2. 110039.

GOODWIN, M.; GOODWIN, C. Emotion within situated activity. **Communication**: An arena of development, v. 33, 2000. p. 33-54.

GROß, A.; DIX, C.; RUUSUVUORI, J.; PERÄKYLÄ; A. Facial gestures in social interaction: Introduction to the special issue. **Social Interaction. Video-Based Studies of Human Sociality**, v. 6, p. 2-11, 2023. DOI: https://doi.org/10.7146/si.v6i3.142894.

HAYASHI, M.; RAYMOND, G.; SIDNELL, J. (org.). **Conversational repair and human understanding**. Cambridge: Cambridge University Press, 2013.

HEATH, C.; HINDMARSH, J.; LUFF, P. **Video in qualitative research**. London: Sage, 2010.

HJULSTAD, J. Practices of organizing built space in videoconference-mediated interactions. **Research on Language and Social Interaction**, v. 49, n. 4, p. 325-341, 2016.

JEFFERSON, G. Glossary of transcript symbols with an introduction. *In*: LERNER, G. H. (org.). **Conversation analysis**. Studies from the first generation. Amsterdam, Philadelphia: John Benjamins, 2004. p. 13-31.

KAUKOMAA, T.; PERÄKYLÄ, A.; RUUSUVUORI, J. Turn-opening smiles: Facial expression constructing emotional transition in conversation. **Journal of Pragmatics**, v. 55, p. 21-42, 2013.

KENDON, A. Some functions of gaze-direction in social interaction. **Acta Psychologica,** v. 26, n. 1, p. 22-63, 1967. DOI: https://doi.org/10.1016/0001-6918(67)90005-4

KENDON, A. **Gesture:** Visible action as utterance. Cambridge: Cambridge University Press, 2004.

LICOPPE, C.; MOREL, J. Video-in-interaction: "Talking heads" and the multimodal organization of mobile and Skype video calls. **Research on Language & Social Interaction**, v. 45, n. 4, p. 399-429, 2012.

LIU, S.; KINGINGER, C. The sociocultural ontogenesis of international students" use of pragmatic strategies in ELF academic communication: Two contrasting case studies. **Journal of Pragmatics**, v. 186, p. 364-381, 2021.

LOENHOFF, J. Interactive technologies and the function of the senses. *In*: NORRIS, S. (org.). **Multimodality in practice**. Investigating theory-in-practice-through-methodology. New York, London: Routledge, 2012. p. 20-35.

LYONS, A. Multimodality. *In*: ZHU, H. (org.). **Research Methods in Intercultural Communication**: A Practical Guide. Malden, Oxford: John Wiley & Sons, 2016. p. 268-280.

MANSTEAD, A. S.; LEA, M.; GOH, J. Facing the future: Emotion communication and the presence of others in the age of video-mediated communication. *In*: KAPPAS, A.; KRÄMER, N. C. (org.). **Face-to-face communication over the internet**. Emotions in a web of culture, language and technology. Cambridge: Cambridge University Press, 2011. p. 144-175.

MONDADA, L. Conversation analysis: Talk and bodily resources for the organization of social interaction. *In*: MÜLLER, C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (org.). **Body – language – communication**. An international handbook on multimodality in human interaction. Volume 1, Berlin, Boston: De Gruyter Mouton, 2013. p. 218-227.

MONDADA, L. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. **Journal of Pragmatics**, v. 145, p. 47-62, 2019. DOI: https://doi.org/10.1016/j.pragma.2019.01.016.

MONDADA, L. Transcription in linguistics. *In*: LITOSSELITI, L. (org.). **Research methods in linguistics**. London: Bloombury, 2018. p. 85-114.

NEVILE, M. The embodied turn in research on language and social interaction. **Research on Language and Social Interaction**, v. 48, n. 2, p. 121-151, 2015.

NORRIS, S.; PRINI, J. Communicating knowledge, getting attention, and negotiating disagreement via video conferencing technology: A multimodal analysis. **Journal of Organizational Knowledge Communication**, v. 3, n. 1, p. 23-48, 2016.

SALOMÃO, A. C. B.; FREIRE JUNIOR, J. C. (org.). **Perspectivas de Internacionalização em casa:** intercâmbio virtual por meio do Programa BRaVE / UNESP. São Paulo: Cultura Acadêmica, 2020.

SCHMIDT, T.; WÖRNER, K. EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. **Pragmatics**, v. 19, p. 565-582, 2009.

SCHMIDT, T.; WÖRNER, K. EXMARaLDA. *In:* MÜLLER, C.; CIENKI, A.; FRICKE, E.; LADEWIG, S. H.; MCNEILL, D.; TEßENDORF, S. (ed.). **Body – Language – Communication: An International Handbook on Multimodality in Human Interaction**. Berlin/Boston: De Gruyter Mouton, 2014. v. 1, p. 402-418.

SCHRÖDER, U. **Co-constructing intercultural space**: An embodied approach [Mouton Series in Pragmatics]. Berlin, New York: De Gruyter, 2025.

SCHRÖDER, U.; NASCIMENTO, T. S. N.; SILVA, A. R. A. Reflexões metodológicas sobre transcrição e a escolha de GAT 2 como sistema de transcrição para o NUCOI. *In*: SCHRÖDER, U.; CARNEIRO MENDES, M. (org.). **Comunicação (inter)cultural em interação**. Belo Horizonte: Editora UFMG, 2019. p. 115-153.

SCHRÖDER, U.; CARNEIRO MENDES, M.; PIRES, C. C.; ALVES DA SILVA; D. H.; NASCIMENTO, T. C.; FIDELIS, F. P. com revisão técnica de P. C. GAGO (UFJF/UFRJ). GAT 2. Um sistema para transcrever a fala-em-interação, traduzido e adaptado do original da Selting, Margret *et al*. **Veredas**, v. 20, n. 2, p. 6-61, 2016.

STIVERS, T. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. **Research on Language and Social Interaction**, v. 41, n. 1, p. 31-57, 2008.

STIVERS, T.; TIMMERMANS, S. Always look on the bright side of life: making bad news bivalent. **Research on Language and Social Interaction**, v. 50, n. 4, p. 404-418, 2017.

STREECK, J. The emancipation of gestures. **Interactional Linguistics**, v. 1, n. 1, p. 90-122, 2021.

SUTTON, V. **The SignWriting Alphabet**. Read and Write any Sign Language in the world. ISWA Manual 2010. The SignWriting Press. Disponível em: http://www.movementwriting.org/symbolbank/. Acesso em: 17 ago. 2024.

Recebido em 7 de setembro de 2024

Aprovado em 18 de outubro de 2024

Anexo – Resumo das convenções de transcrição de GAT 2

SEQUENTIAL STRUCTURE	
.	falling to low final pitch movement of IU
ˆSO	rising-falling accent pitch movement
ˇSO	falling-rising accent pitch movement
´SO	rising accent pitch movement
`SO	falling accent pitch movement
↑	pitch upstep
↓	pitch downstep
<<l>>	lower pitch register
<<h>>	higher pitch register
LOUDNESS AND TEMPO CHANGES	
<<f>>	forte, loud
<<p>>	piano, soft
<<all>>	allegro, fast
<<len>>	lento, slow
<<cresc>>	crescendo, increasingly louder
<<dim>>	diminuendo, increasingly softer
<<acc>>	accelerando, increasingly faster
<<rall>>	rallentando, increasingly slower
CHANGES IN VOICE QUALITY	
<<creaky>>	glottalized
<<whispery>>	change in voice quality as stated
DESCRIPTION OF FURTHER VOCAL AND NON-VOCAL ACTIONS AND OTHER CONVENTIONS	
((laughs))	non-verbal vocal actions and events
<<crying>>	non-verbal vocal actions and events with scope of accompanying speech
(may i)	assumed wording
(i say/let"s say)	possible alternatives
(xxx xxx)	two unintelligible syllables

SEQUENTIAL STRUCTURE	
.	falling to low final pitch movement of IU
^SO	rising-falling accent pitch movement
ˇSO	falling-rising accent pitch movement
´SO	rising accent pitch movement
`SO	falling accent pitch movement
↑	pitch upstep
↓	pitch downstep
<<l>>	lower pitch register
<<h>>	higher pitch register
LOUDNESS AND TEMPO CHANGES	
<<f>>	forte, loud
<<p>>	piano, soft
<<all>>	allegro, fast
<<len>>	lento, slow
<<cresc>>	crescendo, increasingly louder
<<dim>>	diminuendo, increasingly softer
<<acc>>	accelerando, increasingly faster
<<rall>>	rallentando, increasingly slower
CHANGES IN VOICE QUALITY	
<<creaky>>	glottalized
<<whispery>>	change in voice quality as stated
DESCRIPTION OF FURTHER VOCAL AND NON-VOCAL ACTIONS AND OTHER CONVENTIONS	
((laughs))	non-verbal vocal actions and events
<<crying>>	non-verbal vocal actions and events with scope of accompanying speech
(may i)	assumed wording
(i say/let"s say)	possible alternatives
(xxx xxx)	two unintelligible syllables

Fonte: Schröder (2025, p. 103-104)