A temporal and pragmatic analysis of gesture-speech association : a corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system

Patrick Rohrer

DOCTORAT EDUCATION
BRETAGNE LANGAGES, INTERACTIONS
LOIRE COGNITION, CLINIQUE

Nantes Université

upf. Universitat Pompeu Fabra Barcelona

THESE DE DOCTORAT DE

NANTES UNIVERSITE

ECOLE DOCTORALE N° 603
Education, Langages, Interaction, Cognition, Clinique
Spécialité : « *Sciences du Langage* »

UNIVERSITAT POMPEU FABRA

DEPARTAMENT DE TRADUCCIÓ I CIÈNCIES DEL LLENGUATGE

Par

Patrick Louis ROHRER

A temporal and pragmatic analysis of gesture-speech association

A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system

Une analyse temporelle et pragmatique de l'association geste-parole

Une approche basée sur un corpus utilisant le nouveau système d'annotation MultiModal MultiDimensionnel (M3D)

Thèse présentée et soutenue à Barcelone, le 16 décembre 2022

Unité de recherche : Grup d'Estudis de Prosòdia (GrEP) &
UMR 6310 – LLING (Laboratoire de Linguistique de Nantes)

Rapporteurs avant soutenance :

Corine ASTESANO	MCF-HDR, Université Toulouse Jean Jaurès
Stefan BAUMANN	Professeur, Universität zu Köln

Composition du Jury :

Président du jury :	Stefan BAUMANN	Professeur, Universität zu Köln
Examinateurs :	Corine ASTESANO	MCF-HDR, Université Toulouse Jean Jaurès
	Gilbert AMBRAZAITIS	Senior Lecturer, Linnaeus University
Dirs. de thèse :	Pilar PRIETO /	Chercheuse ICREA, Universitat Pompeu Fabra
	Elisabeth DELAIS-ROUSSARIE	Directrice de Recherches CNRS, Nantes Université

To my family, both near and far

Acknowledgements

It's been a long and winding journey to completing my Ph.D. thesis and would have not been possible without the support that I lovingly received from many people.

First, I would like to thank my two thesis supervisors, Pilar Prieto and Elisabeth Delais-Roussarie. Specifically, I would like to thank you, Elisabeth, for your positive and enthusiastic response when I proposed working under your supervision, and your encouragement for doing a cotutelle. You have always remained very interested and flexible through all of the changes over the years (from going between Paris, Nantes, and Barcelona, to all of the ways in which my thesis has evolved) and for that I am truly grateful. Supervising a Ph.D. from another country is not an easy task, but I always felt I could approach you with any of my various (administrative, academic) needs. Finally, I've always felt like I can try new things, come up with new proposals, and your willingness to take them on with me has led me to new ways of thinking, new avenues of research, and the feeling of empowerment and autonomy. For all of this, I am truly grateful to have had you as a Ph.D. Supervisor. Je te remercie énormément !

I would like to thank Pilar for lovingly welcoming me to the GrEP research group. I am immensely grateful that you gave me the opportunity to stay in Barcelona with a Ph.D. grant, which has allowed me to feel like I am truly a part of the research community.

I look fondly back on our regular Monday meetings (and still look forward to the ones to come!) which have always been invigorating and motivating for me, as well as our group lunches with the *Galetes Birba* to wrap up with a coffee and a nice, lively conversation. I cannot emphasize enough how your approach to research and attention to detail have offered me new perspectives, new ideas, and have overall made me a better researcher (and hopefully a better writer and salesman!). Most importantly, your caring nature has made me feel supported not only in an academic sense, but also in a personal sense. You really make GrEP a family, and I am highly honored to be a part of it. Moltíssimes gràcies!

To the both of you, I am highly indebted for the enormous amount of dedication and work you have put into my Ph.D. project, the amount of trust you have confided in me to run projects, and the various opportunities you have given me. I am truly grateful.

I am also thankful to the members of the jury, Gilbert Ambrazaitis, Corine Astésano, and Stefan Baumann, as well as to the alternate members, Frank Kügler and Leo Wanner for taking interest and dedicating the time to review and offer your insights on my work. I am truly honored to have you on my jury.

A special thank you goes to all of the members of the Group of Prosodic Studies who I have had the opportunity to cross paths with. Living and working in Barcelona, we have shared many experiences together and each and every one of you have had an

impact in my life. Iris, Olga, Evi, Florence, Ingrid, Alice, Marusia, Laura - you all were there when I arrived and made me feel incredibly welcome not only to GrEP but also Barcelona. Florence and Marusia, you guys are the best for a night to destress, from the terraces, correfocs, and house parties, thank you for always being around and ready for fun! Ingrid, thank you for being a wonderful colleague to collaborate with, through all the meetings, manuscript revisions, and discussions. More importantly, thank you for your constant support throughout the Ph.D. process, your thoughtfulness (both in and out of the office) and for being a model to follow. Ïo, your smile, and positive energy always boosted my day - and your messages to check in on me during the writing process really did wonders. Thank you so much and keep dancing! Júlia, collaborating with you was such a pleasure, I am greatly inspired by your drive, your intelligence, and your fun-loving nature. And thank you for teaching me how to manage long hair! Yuan, you are one of the kindest people I know and I'm very thankful to have met you. Xiatotong, Peng, thank you both so much for letting me ask you questions about statistics, R, and other conceptual and technical issues and your more than willingness to help me out – but more importantly, thanks for also doing it always with a good laugh. Sara, thanks for your willingness to engage in lively discussions, and for being a reliable and overall wonderful person to work with. To the newer GrEP members, Celia and Ting, thank you both for all of your support in the seminars, and your friendship outside of the office. I would also like to thank the past members of GrEP – Núria, Santi, Alfonso, Paolo, and Maria del Mar. When we get together,

it’s like seeing your cousins, always catching up, having a good laugh. Thank you GrEPpers for being part of this journey!!

I am especially indebted to Ulya Tütüncübasi, the research assistant with whom I closely collaborated for the development of M3D and the English M3D-TED corpus. Our nearly daily, hours-long sessions to talk about gesture, prosody, information structure, ELAN, Praat, and especially metaphors were very enriching for me. I really enjoyed the experience. Thank you for working so hard on the English M3D-TED database, as a lot of the work in this thesis is largely thanks to your contribution to the project. I’m really excited to continue working together, as you always bring something interesting to the table! Thank you!

This thesis would not have been possible without the support that I received from the Department of Translation and Language Sciences, co-financed by the Generalitat de Catalunya. I am deeply grateful for the grant that they awarded me, the funding for a 3-month research stay in Nantes, and for the numerous stipends which have allowed me to travel for conferences from which I benefited immensely. Additionally, I would like to thank the members of the Secretaria, who have always been helpful in resolving the numerous administrative tasks, particularly those that come with being a researcher from abroad.

Working within the joint cotutelle, I had the wonderful opportunity to also be integrated at the Laboratoire de Linguistique de Nantes. I

would like to thank both the laboratoire and the Ecole Doctoral ELICC for their financial support which allowed me to attend multiple conferences and publish an article in open access. Thank you to all of my colleagues, and especially Olivier Crouzet, David Imbert and Gaëlle Ferré for the interest, support, and feedback on my work. Thank you to Sabrina Bendjaballah, Anamaria Falaus, Hamida Demirdache, and Marta Donazzan for your willingness to help me both financially and administratively, and to Monique Loquet for all of the work you do behind the scenes. I would also like to thank the current and previous postdocs, doctoral students, and master's students that I had the pleasure to encounter at the lab: Emmanuella, Kryzzya, Elizabeth, Samantha, Agnieszka, Oana, Lucie, Antoine, Anton, Manon, Ioanna, Amazigh, Pascal, and Jue. Merci beaucoup à tous pour votre précieux soutien, et pour les bonnes rigolades !

A special thank you goes to all of my collaborators for the M3D project – thank you Stefanie Shattuck-Hufnagel, Ada Ren, Núria Esteve-Gibert, Ingrid, Júlia, and Pilar for the wonderful meetings and exchanges we have had over the years. It's always a good sign when a meeting can seem to last for hours because there are so many interesting things to talk about. Also, thank you to Marta Vilà for her help as an M3D labeler! I would also like to thank Aliyah Morgenstern, who introduced me to the field of gesture studies, and who's words at my master's thesis defense encouraged me to take on the challenge of doing a Ph.D. I would also like to thank Marion Tellier and Céline Horgues for agreeing to be members of my CSI,

dedicating the time every year to meet, discuss my Ph.D. project, and for giving me valuable feedback and motivation to continue the project. I would also like to thank Mireia Farrús, Salva Soto-Faraco, and Núria Esteve-Gibert for serving on my jury at the research plan defense, and who also gave me wonderful feedback and ideas for future work. Particularly, thank you Salva and all of the members of the CBC lab for letting me come to your lab and learn about EEG. Also thank you to Patrizia Paggio and all of the GeHM members for giving me the opportunity to present my work, receive valuable feedback, and incorporating me and my work in various projects.

Finally, I am truly grateful for all of the support I have received from my family. Thank you, Mom and Marco, for everything that you do for me, and for your endless love and support from afar. Thank you, Dan, Vane, Liam and Lucy, for always believing in and supporting me. Gracias a mi familia argentina por siempre estar al pendiente de mi tesis y su apoyo desde lejos. Et pour terminer, le plus grand merci à mon compagnon de vie, mon meilleur ami, Sebastian, qui m'a toujours encouragé avec le doctorat, qui m'a aidé, qui m'a motivé et sans qui je n'aurais jamais pu terminer un tel ouvrage. Merci de m'avoir supporté quand j'en avais besoin (surtout pendant les derniers mois de rédaction !). Je t'aime énormément (et Ludito aussi) !

Abstract

Human language is essentially multimodal in that speakers use multiple channels to convey meaning, including speech prosody and gesture (e.g. Mondada, 2016; Perniss, 2018). In the last decades, studies within the field of gesture research have shown both the strong temporal relationship between manual co-speech gestures and prosodic prominence, and have given initial evidence of the relevant pragmatic role of gestures. However, gesture studies have shown a tendency to focus on the role of prosodic prominence alone as the main attractor for gesture production, and little empirical research has systematically assessed the role of prosodic phrasal structure in the attraction of gesture, or the joint contribution of gestural and prosodic prominence for pragmatic effects, particularly in terms of signaling information structure (henceforth, IS). Furthermore, no studies have specifically accounted for potential difference in gesture type (i.e., referential vs. non-referential gestures). In our view, a multidimensional analysis of independent aspects of gesture is crucial to allow for a systematic assessment of their different prosodic and pragmatic characteristics. The two main goals of this thesis will be to develop a novel gesture labeling system (i.e., the MultiModal MultiDimensional (M3D) system) and to apply the system to better understand the prosodic and pragmatic characteristics of both referential and non-referential gestures.

The present Ph.D. thesis consists of four independent studies plus introductory and conclusion sections that unite the four studies. The first study proposes M3D as a novel tool for multidimensional

gesture annotation that is in line with the advancing theories in the field of gesture studies. Through the application of M3D to a corpus of French TED Talks (five TED Talks with over 37 minutes of multimodal speech), the second study shows how phrase-initial accents act as strong gestural attractors regardless of gesture type, and how the production of multiple subsequent gestures is largely guided by the temporal duration of prosodic phrases. To further examine the effects of phrasal position, a third study was carried out on English TED Talks (five TED Talks with over 28 minutes of multimodal speech), assessing the temporal association of gestures with pitch accentuation while systematically taking into account the effects of nuclear status and degrees of relative prominence. The results highlight the role of prenuclear pitch accentuation as a strong attractor of gesture, independent of relative prominence. Finally, the fourth study assesses the joint role of prosody and gesture in the marking of IS (particularly, the information status of referents; henceforth, ISR) in the same corpus of English TED Talks. The results show how prominence (via pitch accentuation) and the production of gesture work together to mark newer information in speech, with pitch accent type and gesture type not playing key roles as cues to ISR.

All in all, the four studies contained in this thesis offer a novel gesture annotation tool that can be used for the development of multimodal corpora accounting for a variety of aspects of speech, gesture, and prosody. The empirical studies further our knowledge about the temporal association of gesture and speech, showing that not only prosodic prominence, but also prosodic phrasing are key to

understanding the relationship between the two channels. The studies also further our knowledge in terms of how these two channels interact to convey pragmatic meaning. Thus, this multidimensional analysis of gesture greatly contributes to the ongoing effort to elucidate the precise nature of the temporal and pragmatic properties of both referential and non-referential gestures in discursive speech.

Resum

El llenguatge humà és per naturalesa multimodal, ja que els parlants utilitzen múltiples mitjans, com ara la prosòdia i el gest, per transmetre significats comunicatius (p. ex., Mondada, 2016; Perniss, 2018). En les últimes dècades, estudis dins de l'àmbit de la investigació gestual han demostrat la forta relació temporal entre els gestos manuals i la prominència prosòdica, i han començat a mostrar el rol pragmàtic dels gestos. Tanmateix, aquests estudis s'han centrat en el paper de la prominència prosòdica com a principal pol d'atracció per la producció de gestos, i poca investigació empírica ha avaluat sistemàticament el paper de l'estructura prosòdica de la frase en aquest procés. També se sap poc sobre els efectes pragmàtics de la gestualitat, especialment pel que fa a l'estructura de la informació (d'ara endavant, IS). A més a més, cap estudi no ha tingut en compte la possible diferència entre els trets temporals i pragmàtics dels diferents tipus de gest (és a dir, entre els gestos referencials i no referencials). Al nostre parer, una anàlisi multidimensional del gest és crucial per permetre una avaluació sistemàtica de les seves característiques prosòdiques i pragmàtiques. Els dos objectius principals d'aquesta tesi seran desenvolupar un nou sistema d'etiquetatge gestual (és a dir, el sistema MultiModal MultiDimensional (M3D)) i aplicar el sistema per entendre millor les característiques prosòdiques i pragmàtiques dels gestos tant referencials com no referencials.

La present tesi doctoral consta de quatre estudis independents a més a més de les seccions d'introducció i conclusions que uneixen els

quatre estudis. El primer estudi proposa el sistema M3D com una nova eina per a l'anotació multidimensional de gestos que està en línia amb les teories més avançades del camp. Mitjançant l'aplicació de l'M3D a un corpus de TED Talks en llengua francesa (cinc TED Talks amb més de 37 minuts de parla multimodal), el segon estudi mostra com els accents tonals que es troben a inici de la frase actuen com a fort punt d'ancoratge per a la gestualitat, independentment del tipus de gest, i com la producció de múltiples gestos contigus es guia en gran part per la durada temporal de les frases prosòdiques. Per examinar més a fons els efectes de la posició de la frase, es va dur a terme un tercer estudi sobre un corpus de TED Talks en llengua anglesa (cinc TED Talks amb més de 28 minuts de parla multimodal). L'estudi va avaluar l'associació temporal dels gestos amb l'accentuació prosòdica tenint en compte de manera sistemàtica els efectes de nuclearitat dels accents i del seu grau de prominència. Els resultats destaquen el paper de l'accentuació prenuclear com a fort pol d'atracció del gest, independentment de la seva prominència relativa. Finalment, el quart estudi avalua el paper conjunt de la prosòdia i el gest en el marcatge de l'IS (en particular, l'estat informatiu dels referents; d'ara endavant, ISR pel nom en anglès) en el mateix corpus de TED Talks anglesos. Els resultats mostren com la prominència (mitjançant l'accentuació prosòdica) i la producció gestual funcionen junts per marcar la informació més nova del discurs. Tanmateix, el tipus d'accent tonal i el tipus de gest no juguen un paper clau com a marcadors de l'ISR.

En resum, els quatre estudis presentats en aquesta tesi ofereixen una nova eina d'anotació gestual que es pot utilitzar per al desenvolupament de corpus multimodals que tenen en compte diversos aspectes de la parla, el gest i la prosòdia. Els resultats dels estudis empírics amplien el nostre coneixement sobre l'associació temporal entre el gest i la parla i demostren que no només la prominència prosòdica, sinó també el fraseig és una peça clau per entendre la relació temporal entre gest i parla. Els estudis també milloren el nostre coneixement sobre com aquests dos canals interactuen per transmetre significats pragmàtics com l'estructura informativa. Així, aquesta anàlisi multidimensional del gest contribueix en gran mesura a l'esforç actual per dilucidar de forma més precisa la naturalesa de les propietats temporals i pragmàtiques dels gestos referencials i no referencials en el discurs.

Resumen

El lenguaje humano es por naturaleza multimodal en el sentido de que los hablantes utilizan múltiples canales para transmitir el significado, incluyendo la prosodia y los gestos (por ejemplo, Mondada, 2016; Perniss, 2018). En las últimas décadas, estudios en el ámbito de la investigación gestual han demostrado la fuerte relación temporal entre los gestos manuales y la prominencia prosódica y han empezado a mostrar el rol pragmático de los gestos. Sin embargo, estos estudios se han centrado en el papel de la prominencia prosódica como principal polo de atracción para la producción de gestos y poca investigación empírica ha evaluado sistemáticamente el papel de la estructura prosódica de la frase en este proceso. También se sabe poco sobre los efectos pragmáticos de la gestualidad, especialmente por lo que se refiere a la estructura de la información (en adelante, IS). Además, ningún estudio ha tenido en cuenta específicamente la diferencia potencial en las características temporales y pragmáticas de los diferentes tipos de gestos (es decir, entre los gestos referenciales frente a los no referenciales). En nuestra opinión, un análisis multidimensional del gesto es crucial para permitir una evaluación sistemática de sus características prosódicas y pragmáticas. Los dos objetivos principales de esta tesis serán desarrollar un novedoso sistema de anotación gestual (es decir, el sistema MultiModal MultiDimensional (M3D) y aplicar ese sistema para comprender mejor las características prosódicas y pragmáticas de los gestos referenciales y no referenciales.

La presente tesis doctoral consta de cuatro estudios independientes y las secciones de introducción y conclusión que sirven de unión entre ellos. El primer estudio propone el sistema M3D como una herramienta novedosa para la anotación de gestos multidimensionales en consonancia con las teorías más avanzadas en el ámbito. Mediante la aplicación de M3D a un corpus de TED Talks en lengua francesa (cinco TED Talks con más de 37 minutos de discurso multimodal), el segundo estudio muestra cómo los acentos tonales que se encuentran al inicio de la frase actúan como fuertes atractores gestuales, independientemente del tipo de gesto, y cómo la producción de múltiples gestos contiguos sirve de guía en gran medida por la duración temporal de las frases prosódicas. Para examinar más a fondo los efectos de la posición de la frase, se llevó a cabo un tercer estudio sobre TED Talks en lengua inglesa (cinco TED Talks con más de 28 minutos de discurso multimodal). El estudio evalúa la asociación temporal de los gestos con la acentuación prosódica, teniendo en cuenta de manera sistemática los efectos del estado nuclearidad de los acentos y del grado de prominencia. Los resultados ponen de relieve el papel de la acentuación prenuclear como un polo de atracción del gesto, independientemente de su prominencia relativa. Por último, el cuarto estudio evalúa el papel conjunto de la prosodia y el gesto en el marcaje de la IS (en particular, el estado informativo de los referentes; en adelante, ISR) en el mismo corpus de TED Talks en inglés. Los resultados exponen cómo la prominencia (a través de la acentuación prosódica) y la producción gestual actúan conjuntamente para marcar la información más nueva en el

discurso. Sin embargo, ni el tipo de acento tonal ni el tipo de gesto juegan un papel clave como señales de la ISR.

Resumiendo, los cuatro estudios presentados en esta tesis ofrecen una novedosa herramienta de anotación gestual que se puede utilizar para el desarrollo de corpus multimodales que den cuenta de diversos aspectos del habla, de los gestos y de la prosodia. Los resultados de los estudios empíricos amplían nuestros conocimientos sobre la asociación temporal entre los gestos y el habla, y demuestran que no sólo la prominencia prosódica, sino también el fraseo es una pieza clave para entender la relación temporal entre gesto y prosodia. Los estudios también amplían nuestros conocimientos en cuanto a la forma en que estos dos canales interactúan para transmitir significados pragmáticos como la estructura informativa. Así pues, el análisis multidimensional del gesto presentado contribuye en gran medida al esfuerzo actual por dilucidar de forma más precisa la naturaleza de las propiedades temporales y pragmáticas de los gestos referenciales y no referenciales en el discurso.

Sommaire

Pour communiquer, les êtres humains font appel à différentes stratégies afin que leur(s) interlocuteur(s) puisse(nt) comprendre le plus précisément et facilement possible le sens du message qu'ils souhaitent transmettre. Dans la communication orale, la signification d'un message se calcule (ou se construit) non seulement à partir de la chaîne de sons associés aux mots et à partir de la façon dont ces derniers sont structurés en fonction des règles morphosyntaxiques, mais aussi grâce à des éléments suprasegmentaux intonatifs et rythmiques qui se superposent, sur le plan sonore, aux unités segmentales. De plus, le corps du locuteur est un élément central pour la transmission du message et l'interprétation correcte de l'énoncé. Il fournit en effet des informations complémentaires. Dans la présente thèse, nous soutenons une vision globale du langage qui englobe l'utilisation de plusieurs modes de communication dans la construction du sens, notamment la prosodie de la parole, le regard, les gestes manuels, les mouvements de la tête, les expressions faciales, les postures corporelles, etc. (voir, par exemple, Goodwin, 2000 ; Mondada, 2016 ; Perniss, 2018). Plus précisément, la présente thèse se concentre sur les *gestes co-verbaux* et ses interactions avec la structure prosodique du langage. Selon la définition de Kendon (2004, p. 7), le geste fait référence à « une action visible de n'importe quelle partie du corps, lorsqu'elle est utilisée en tant qu'énonciation, ou en tant que partie d'une énonciation ». « Trois règles de synchronie » ont été proposés (McNeill, 1992) et constituent l'un des principaux arguments en faveur de la

conceptualisation du geste comme élément clé du langage. Selon ces règles, le geste et la parole sont produits en même temps (*règle de synchronie phonologique*), sont cohérents sur le plan sémantique (*règle de synchronie sémantique*) et sont cohérents sur le plan pragmatique (*règle de synchronie pragmatique*).

La règle de synchronie phonologique a été avancée à partir des observations initiales de Kendon (1980) selon lesquelles la partie du mouvement gestuel qui donne le sens de celui-ci, autrement appelé « stroke », précède ou s'achève avec, mais pas après le noyau des syllabes prominents (« phonological peak syllable » de l'énoncé (McNeill, 1992, p. 26). Lorsque cette règle est menacée, il a été démontré que les locuteurs cessent temporairement de produire des gestes afin de maintenir la synchronie phonologique, comme par exemple dans le cas de disfluence (Graziano & Gullberg, 2018). Selon la règle de la synchronie sémantique, la parole et le geste devraient refléter la même idée, puisqu'ils sont conceptualisés ensemble. Ainsi, le geste et la parole expriment conjointement le même sens central tout en décrivant différents aspects de celui-ci (McNeill, 2000, p. 7). Enfin, McNeill (2000) propose la règle de synchronisation pragmatique selon laquelle le geste et la parole partagent la même fonction pragmatique. Dans l'ensemble, la règle de synchronisation pragmatique indique que les deux modes de communication fonctionnent ensemble pour introduire le genre du discours à venir.

De nombreuses propositions ont été avancées dans le but de classifier les gestes, le système de McNeill (1992) étant le plus

utilisé dans le domaine de la recherche en gestuelle aujourd'hui. McNeill (1992) fait la distinction entre quatre principaux types de gestes : les gestes iconiques, métaphoriques, déictiques et les gestes de battement. Les gestes iconiques représentent de manière imagée une action ou un objet concret, en lien étroit avec le sens sémantique du discours qu'ils accompagnent. Comme les gestes iconiques, les gestes métaphoriques véhiculent une représentation imagée, mais au lieu d'actions ou objets concrets, ils représentent des concepts ou des idées abstraites. Par exemple, prononcer « le voyage prendra trois jours » tout en produisant un geste avec une main se déplaçant vers la droite, représentant le temps (le concept abstrait de trois jours) sur un axe physique horizontal (voir également Cienki & Müller, 2008). Les gestes déictiques font référence aux gestes de pointage désignant des entités dans l'espace. La dernière catégorie de gestes décrite par McNeill est celle des gestes de battement. Ces gestes ne représentent pas de contenu sémantique et ne font pas référence à des entités dans l'espace. Traditionnellement, ces gestes ont été classés sur la base de leur forme, comme étant de simples mouvements de la main ou du doigt s'associant à une proéminence prosodique et qui semblent « marquer un rythme musical » (McNeill, 1992, p. 15). Par ailleurs, une valeur pragmatique-discursive a été attribuée aux gestes de battement, dans la mesure où ils soulignent l'importance pour le discours des mots ou les phrases qu'ils accompagnent (McNeill, 1992, p. 15).

Cependant, si l'on s'accorde sur la théorie de McNeill selon laquelle tous les gestes, quel que soit leur type, présentent une synchronisation sémantique, pragmatique et phonologique avec la

parole, il ne semble pas correct de mettre en évidence le rôle prosodique et pragmatique des seuls gestes de battement, et de diviser les autres gestes en se basant uniquement sur leurs propriétés sémantiques. Au lieu de types de gestes, McNeill (2006) suggère d'adopter le terme de « dimensions » de signification, où les gestes peuvent illustrer différents niveaux d' « iconicité », de « métaphoricité », de « deixis » ou de « marquage temporel ». Dans cette perspective, des ouvrages théoriques récents ont commencé à soutenir une approche dimensionnelle de l'étude des gestes (Prieto et al., 2017 ; Shattuck-Hufnagel & Prieto, 2019). Une telle approche met au premier plan les trois règles de synchronie de McNeill en reconnaissant que les propriétés sémantiques, pragmatiques et prosodiques des gestes doivent être évaluées de manière indépendante. Dans la pratique, cette proposition implique que les chercheurs évaluent systématiquement la cohérence sémantique des gestes avec le discours (s'ils sont référentiels - c'est-à-dire qu'ils renvoient à un contenu sémantique via l'iconicité, la métaphoricité et la deixis - ou non référentiels), leur cohérence pragmatique avec le discours (s'ils contribuent au sens pragmatique), et leur co-occurence temporelle avec le discours (s'ils sont produits en synchronie avec la structure prosodique).

Cette thèse poursuit un double objectif. Tout d'abord, elle propose une nouvelle approche de l'annotation des gestes co-verbaux qui épouse une vision dimensionnelle, selon laquelle les chercheurs devraient considérer les caractéristiques sémantiques, pragmatiques et prosodiques des gestes d'une manière non mutuellement exclusive. En second lieu, cette thèse vise à mieux comprendre les

relations prosodiques et pragmatiques des gestes référentiels et non référentiels, en particulier la façon dont la structure prosodique complexe influence les modèles de production gestuelle, et comment ces deux modes de communication interagissent pour des raisons pragmatiques. Pour tenter de répondre à ces objectifs, le corps de cette thèse s'organise autour de quatre études individuelles.

L'étude du **Chapitre 2** a deux objectifs : (a) recenser les caractéristiques des dix principaux systèmes d'annotation multimodale actuellement disponibles ; et (b) décrire la structure tri-dimensionnelle du système d'annotation M3D, un système d'étiquetage proposé en accès libre comprenant un ensemble de conventions d'annotation fiables, des supports de formation et un corpus audiovisuel d'une heure étiqueté. En ce qui concerne le premier objectif, il s'agit d'évaluer les caractéristiques communes et plus standardisées de l'annotation multimodale et les principales différences entre les systèmes. Les résultats de cette analyse montrent que la communauté des chercheurs en gestuelle n'a pas encore mis au point un système d'annotation qui fasse l'objet d'un consensus général. En particulier, la structure d'analyse de la signification des gestes varie considérablement d'un système à l'autre. Il est important de noter qu'aucun des systèmes examinés ne prend en compte le point de vue plus récent selon lequel les gestes ne devraient pas être considérés comme appartenant à des catégories mutuellement exclusives (McNeill, 2006 ; Prieto et al., 2018 ; Shattuck-Hufnagel & Prieto, 2019). En outre, ces systèmes d'annotation n'intègrent pas systématiquement les dimensions des

gestes qui peuvent être superposées (la forme, les caractéristiques prosodiques et les significations sémantiques ou pragmatiques). Si tous les systèmes examinent le langage multimodal en incluant au moins deux modes de communication (la parole et le geste), seuls trois d'entre eux adoptent ouvertement une approche multimodale qui vise à comprendre les interactions entre les modes de communication. De plus, ces systèmes n'évaluent pas toujours de manière approfondie les contributions pragmatiques potentielles du geste. Le système M3D s'appuie sur ces systèmes existants tout en incorporant la transcription indépendante des différents modes de communication dans le langage multimodal (c'est-à-dire, la parole, la prosodie et le geste) grâce à un système d'annotation tri-dimensionnel du geste qui permet d'étudier les caractéristiques gestuelles de manière non mutuellement exclusive.

Selon les trois règles de synchronie établies par McNeill (1992), les comportements gestuels sont intégrés à la proéminence prosodique et transmettent une signification à la fois sémantique et pragmatique, ce qui souligne l'importance de développer une approche qui rende compte de ces multiples dimensions du geste lié à la parole, notamment sa forme cinématique, ses propriétés prosodiques et ses contributions sémantiques et pragmatiques. Le système M3D apporte une contribution majeure en ce qu'il est fondé sur les trois dimensions suivantes: 1) la *dimension de la forme du geste*, qui se réfère à un certain nombre d'aspects physiques du geste à travers différents articulateurs, y compris la configuration et les caractéristiques cinématiques du geste ; 2) la *dimension prosodique*, qui se réfère à l'association des gestes à la structure prosodique via

un ensemble de procédures perceptives standardisées, ainsi que les caractéristiques prosodiques des gestes, y compris leur caractère rythmique et leur structure en phases successives (l'organisation progressive des mouvements) et 3) la *dimension du sens*, qui capture les informations sémantiques (c'est-à-dire la référentialité) et pragmatiques qui peuvent être exprimées dans le geste. En outre, le système M3D intègre une étude plus classique et plus approfondie du geste et offre des ressources pour son application à différentes données. De cette façon, les chercheurs peuvent développer des corpus comparables, éviter une simplification excessive des actes de communication multimodaux complexes, et adopter des méthodes d'étiquetage claires. Le système M3D a été appliqué pour l'étiquetage des corpus M3D-TED et s'est avéré fiable pour le codage d'aspects importants des gestes, notamment leur découpage en phases successives, l'identification des apex (le pic cinématique du « stroke »), la référentialité des gestes et leur fonction pragmatique.

Les **Chapitres 3 et 4** mettent en œuvre le système M3D afin de mieux comprendre la règle de synchronie phonologique, c'est-à-dire comment le geste est associé temporellement à la structure prosodique. De nombreuses études sur l'association temporelle entre le geste et l'accentuation (i.e., « pitch accent ») ont montré que la proéminence dans le geste (c'est-à-dire le stroke ou l'apex du geste) et la proéminence dans la parole (c'est-à-dire les syllabes accentuées) tendent à être produits dans une synchronie temporelle étroite (Loehr, 2004 ; Yasinnik et al., 2004 ; Jannedy & Mendoza-Denton, 2005 ; Leonard & Cummins, 2011 ; Esteve-Gibert & Prieto,

2013 ; Shattuck-Hufnagel & Ren, 2018 ; Pouw & Dixon, 2019b). Cependant, elles ont été menées dans des langues comme l'anglais où l'accentuation est liée à la prominence des termes auxquels elle est associée, qu'il s'agisse de mot lexical, de syntagme, etc. En français, en revanche, l'accentuation, pour un part en tout cas (notamment l'accent final), a une fonction démarcative et n'indique pas la proéminence du mot ou du syntagme à la fin duquel elle est réalisée. Deux études sur le français ont révélé que les gestes coïncident avec l'accentuation prosodique (Ferré, 2014 ; Roustan & Dohen, 2010). Une autre étude a trouvé que les gestes ont tendance à chevaucher plusieurs AP subséquents (Ferré, 2010). À notre connaissance, aucune étude sur le français n'a examiné des phases plus petites au sein du geste (e.g., la phase obligatoire, autrement appelée « stroke », qui est considérée comme la partie significative du geste), et sa relation précise avec l'accentuation.

En outre, les études précédentes sur l'association temporelle entre le geste et l'accentuation ont analysé deux éléments gestuels séparément : le stroke et l'apex. Dans la plupart des études de laboratoire portant sur ce dernier, il a été trouvé un lien étroit entre apex et accent tonal, tandis que de nombreuses études portant sur la parole naturelle ont utilisé des fourchettes temporelles plus grandes pour évaluer leur cooccurrence temporelle. Peu d'études ont directement comparé l'association prosodique de ces deux éléments gestuels en tenant compte de la référentialité du geste. Ainsi, le premier objectif de l'étude du **Chapitre 3** est d'analyser spécifiquement les deux éléments gestuels et leur association avec les accents initiaux (IA) et finaux (FA) dans la parole naturelle en

français (hors-laboratoire), et en comparant les gestes référentiels et non-référentiels.

De plus, on en sait beaucoup moins sur les modèles rythmiques dans la production de gestes consécutifs, appelés groupes rythmiques de gestes (GRG). Les seules études dont nous disposons suggèrent que la production de GRG est en grande partie indépendante du rythme de la parole, un seul geste du groupe étant associé à l'accent nucléaire (Loehr, 2007 ; McClave, 1994). Le deuxième objectif de l'étude est de tester les affirmations selon lesquelles les GRG non référentiels sont plus rythmiques que les GRG référentiels, à la fois en termes de fréquence et d'isochronie au sein du groupe. Enfin, l'étude examine la relation entre le rythme de la parole et la production de GRG.

Ainsi, nous pouvons identifier les trois questions de recherche suivantes :

1.) L'accentuation continue-t-elle à agir comme un ancrage prosodique pour le geste en français, et cette relation est-elle modulée par le type d'accent (IA vs FA) ou le type de geste (référentiel vs non référentiel) ?

2.) Les GRG ont-ils tendance à être plus non référentiels par nature, et sont-ils plus isochrones que les groupes de gestes référentiels en forme de battement ?

3.) Les GRG ont-ils tendance à marquer les AP subséquentes en français, et cette relation est-elle modulée par l'accentuation (c.-à-d. la présence ou l'absence d'IA) ? Si non, cette relation est-elle sensible à la durée des phrases prosodiques ?

Pour répondre à ces questions de recherche, une analyse de corpus a été effectuée sur le corpus M3D-TED français. Le corpus contient plus de 37 minutes de discours multimodal de cinq locuteurs français natifs donnant un TED Talk (durée moyenne : 07m 30s). Un échantillon d'environ 5-10 minutes de discours par locuteur a été choisie pour l'annotation. Pour cet échantillon, les locuteurs ont produit une moyenne de 300,6 gestes (± 87,47), et une moyenne de 779,4 syllabes accentuées (± 234,15).

En ce qui concerne la première question de recherche, nous avons constaté que les strokes s'alignent avec les syllabes accentuées à des degrés similaires à ceux qui ont été précédemment rapportés pour l'anglais (e.g., Shattuck-Hufnagel & Ren, 2018). À notre connaissance, cette étude est la première à examiner les modèles d'alignement en français de manière à être comparables aux études précédentes en anglais. Cependant, l'alignement des apex était beaucoup plus variable par rapport à ce qui a été rapporté dans la littérature précédente. Il est important de noter que l'étude actuelle apporte un éclairage supplémentaire sur les modèles d'alignement, en constatant que lorsque l'IA et le FA sont tous les deux présents dans un AP, le geste s'aligne avec l'IA beaucoup plus souvent

qu’avec le FA. Il est important de noter qu'aucune différence significative n'a été trouvée en ce qui concerne la référentialité de geste, dans la lignée de ce qui a été précédemment rapporté dans la littérature (e.g., Shattuck-Hufnagel & Ren, 2018).

En ce qui concerne la deuxième question de recherche, nous avons constaté que les GRG avaient tendance à se composer entièrement de gestes non référentiels plus souvent que couplés à des gestes référentiels, ce qui étaye l'idée que les gestes non référentiels peuvent se regrouper pour « battre un rythme musical » de la parole, comme le décrit McNeill (1992, p. 15). De plus, l’analyse de l'isochronie des GRG montre que les GRG non référentiels sont plus isochrones que les référentiels.

Enfin, en ce qui concerne la troisième question de recherche, nous avons constaté que si la plupart des AP ne contiennent qu'un seul apex du GRG, une relation biunivoque apex - AP au sein d’un GRG ne s'est pas avérée être une tendance majeure. En d'autres termes, les AP au sein d’un GRG peuvent être omises ou doublement marqués par un apex. Il est important de noter que ce type de marquage gestuel ne semble pas être déterminé par des schémas d'accentuation. En d'autres termes, les AP qui présentent deux accents ne sont pas forcément marqués par deux gestes. Ces résultats sont en accord avec les études précédentes portant sur l'anglais et qui ont suggéré que les apex des GRG ont tendance à être produits indépendamment des modèles d'accentuation (McClave 1994 ; Loehr, 2007). Toutefois, les résultats de la présente étude montrent que la relation entre la structure prosodique

et la production rythmique du geste n'est pas entièrement indépendante, du moins en français. Les résultats du modèle de régression linéaire suggèrent en effet que la durée temporelle des AP prédit significativement l'intervalle de temps entre les apex des GRG. Ainsi, plus les AP s'allongent, plus les distances entre les apex subséquents augmentent.

Les principaux résultats de cette étude montrent pour la première fois que le geste peut avoir une attirance particulière pour le frontière gauche de la phrase prosodique, via leur association avec les IA en français. De plus, elle est la première à montrer que le rythme de la parole et le rythme gestuel ne sont pas complètement indépendants, et que la durée du phrasé prosodique prédit directement l'intervalle entre les apex successifs du GRG. Cette étude contribue à élargir les connaissances sur l'intégration geste-parole dans une langue relativement peu étudiée sur ces questions. De plus, elle contribue à notre compréhension de la production rythmique des gestes subséquents, un sujet qui reste encore à développer.

L'étude du **Chapitre 4** a pour but d'approfondir l'effet de frontière gauche trouvé dans les résultats de l'étude du **Chapitre 3**, notamment en étudiant les modèles d'alignement en anglais, en démêlant spécifiquement le rôle de la position phrastique et du degré relatif de proéminence. En effet, les premières observations sur l'association temporelle entre le geste et l'accentuation en anglais ont suggéré que les gestes s'associent spécifiquement aux accents nucléaires (Kendon, 1980 ; McNeill, 1992). Comme dans le

Chapitre 3, cette étude examine spécifiquement ces deux repères gestuels dans le discours naturel en anglais, en comparant les gestes référentiels et non référentiels. Le deuxième objectif de l'étude est d'évaluer le rôle de la nucléarité (accent pré-nucléaire vs. nucléaire) dans la synchronisation geste-parole et de déterminer si cette relation est motivée par le degré de proéminence relative ou le positionnement phrastique. Ainsi, nous pouvons identifier les trois questions de recherche suivantes :

1.) Les strokes et les apex des gestes s'alignent-ils avec les syllabes accentuées dans les TED Talks en anglais, et cette relation est-elle modulée par le statut référentiel des gestes ?

2.) Les gestes sont-ils associés aux accents nucléaires plus qu'aux accents pré-nucléaires ?

3.) Cette relation est-elle déterminée par les degrés de proéminence relatives ou par la position phrastique ?

Le corpus M3D-TED anglais a été utilisé pour cette étude. Ce corpus contient plus de 23 minutes de discours multimodal annoté. Il est constitué de cinq TED Talks (durée moyenne : 4m 47s) produits par cinq locuteurs natifs anglais. Une moyenne de 277,8 gestes, et de 399 syllabes accentuées par locuteur a été annotée.

Comme pour l'étude précédente (**Chapitre 3)**, les strokes gestuels s'alignent essentiellement avec les syllabes accentuées, tandis que

les apex montrent une variabilité beaucoup plus grande dans leur alignement. Dans les deux cas, cette relation n'est pas affectée par le statut référentiel des gestes. Il est important de noter que nous avons aussi trouvé un effet de « marquage de la frontière gauche ». En d'autres termes, les gestes s'associent aux accents pré-nucléaires au sein d'une phrase intermédiaire, sans que le degré de proéminence relative de ces accents ait un rôle à jouer. De plus, bien que cela ne soit pas statistiquement significatif, le geste tend à marquer le premier accent dans la phrase intermédiaire (ou syntagme intermédiaire *ip*). La contribution principale de cette étude est de démontrer que non seulement la proéminence prosodique, mais aussi la structure prosodique complexe (y compris le phrasé) peuvent être des facteurs clés dans la synchronisation geste-parole.

Enfin, l'étude du **Chapitre 5** examine les indices multimodaux permettant de marquer le statut informationnel des référents dans les TED Talks en anglais. Le statut informationnel des référents (ISR) indique si les entités du discours sont nouvelles, accessibles via le contexte ou la connaissance partagée, ou données (ayant déjà été introduites dans le discours ; voir Götze et al., 2007 ; Krifka, 2008 pour une revue). La plupart des études précédentes sur la question n'ont étudié qu'un seul mode à la fois (en se concentrant soit sur le marquage prosodique de l'ISR, soit sur le marquage gestuel de l'ISR individuellement). Les études sur le marquage prosodique de l'ISR ont constaté que les référents donnés ont tendance à être désaccentués, tandis que les référents nouveaux ou accessibles sont plus susceptibles de recevoir des accents (e.g., Halliday, 1967 ; Chafe, 1974 ; Brown, 1983 ; Gussenhoven, 1984 ; Hirschberg, 1993

; Cruttenden, 1997 ; Hirschberg, 2002 ; Prince 1981 ; Ladd, 1980, 2008 parmi beaucoup d'autres ; voir Kügler & Calhoun, 2020 pour une revue). De la même manière, le geste est plus susceptible d'être produit avec des informations nouvelles qui font avancer le discours (par exemple, Debreslioska et al., 2013 ; Debreslioska & Gullberg, 2019 ; Gullberg, 2003 ; Levy & Fowler, 2000 ; Marslen-Wilson et al., 1982 ; McNeill, 1992 ; Yoshioka, 2008). Bien que de nombreux ouvrages ont mis en évidence le lien étroit entre la prosodie et le geste (e.g., Loehr, 2012), aucune étude à notre connaissance n'a étudié conjointement ces deux indices multimodaux comme marqueur de référents. En effet, la plupart des études se sont concentrées soit sur la présence ou l'absence d'accentuation, soit sur le type d'accent ToBI comme marqueurs de l'ISR. C'est pourquoi l'objectif de cette étude est d'une part d'évaluer les contributions de la proéminence relative ainsi que du type d'accent dans le marquage d l'ISR, et d'autre part d'analyser le marquage gestuel de l'ISR, notamment dans les positions pré-nucléaires accentuées. Les trois questions de recherche suivantes peuvent être formulées :

1.) Comment le geste et l'accentuation marquent-ils conjointement l'ISR dans les TED Talks anglais ?

2.) En termes de prosodie, quelle est la relation entre le degré de proéminence relative, le type d'accent ToBI et l'ISR ?

3.) En termes de gestes, le type de geste (c'est-à-dire référentiel ou non référentiel) joue-t-il un rôle dans

le marquage de l'ISR ? Y a-t-il marquage gestuel de l'ISR dans les positions pré-nucléaires ?

L'annotation de l'ISR a été effectuée sur le même corpus M3D-TED anglais décrit dans les paragraphes précédents (Chapitre 4). Plus spécifiquement, le système d'étiquetage LISA simplifié décrit par Götze et al. (2007) a été appliqué au corpus, dans lequel le statut de l'information est appliqué à l'expression référentielle (c'est-à-dire le NP ou le PP entier). L'analyse de geste comme marqueur d'ISR se base essentiellement sur une association temporelle non stricte (par exemple, Rohrer et al., 2019) et/ou sur le sens sémantique véhiculé par le geste.

En ce qui concerne la première question de recherche, les résultats de l'étude montrent que l'accentuation et le geste marquent conjointement la structure de l'information, les référents nouveaux et accessibles recevant principalement un double marquage par la prosodie et le geste, tandis que les référents donnés reçoivent un marquage multimodal significativement moins important que les autres types de référents. En ce qui concerne la deuxième question de recherche, les résultats de l'étude actuelle n'ont pas trouvé de preuve d'une correspondance univoque entre l'ISR et le type d'accent ou le type de geste. Cependant, le degré de proéminence relative en général semble effectivement être fonction de l'ISR. En ce qui concerne la dernière question de recherche, lorsque les référents sont marqués par des accents pré-nucléaires, les référents accessibles sont plus susceptibles de recevoir un geste. Cette étude apporte une contribution aux connaissances sur le marquage

multimodal de l'ISR en soulignant le rôle de la proéminence relative plutôt que celui du type d'accent ; elle éclaire aussi sur l'importance des gestes, notamment dans des contextes pré-nucléaires.

Dans l'ensemble, les résultats des quatre études de la thèse montrent que le système M3D est un outil utile pour analyser le langage multimodal. En premier lieu, les résultats des chapitres 3, 4, et 5 démontrent que le langage est multimodal par nature, que les locuteurs utilisent de multiples stratégies de construction du sens pour communiquer et, surtout, que ces modes interagissent à différents niveaux, tant en termes de coproduction temporelle que de transmission du sens pragmatique (en relation à la structure informationnelle). En second lieu, les résultats soutiennent la nécessité d'une analyse tri-dimensionnelle des gestes, à savoir leur forme, leurs caractéristiques prosodiques et leur signification (sémantique et/ou pragmatique). La caractérisation traditionnelle des gestes de « battement » comme étant les seuls gestes ayant une fonction de marqueurs prosodiques de proéminence pour les fonctions pragmatiques du discours n'est pas observée dans les deux corpus M3D-TED, français et anglais. En effet, les gestes de tous types, qu'ils soient référentiels ou non référentiels, sont généralement associés à l'accentuation, tant en français qu'en anglais. Il est intéressant de noter toutefois que, bien que les gestes référentiels et non référentiels soient produits de manière rythmique, les résultats suggèrent que les gestes non référentiels sont plus susceptibles d'être perçus comme se produisant de manière consécutive et rythmique. De plus, les gestes ne sont pas *ou* sémantiques *ou* pragmatiques dans leur contribution au sens de la

parole ; nous démontrons en effet que les gestes référentiels et non référentiels signalent l'ISR à la même hauteur. Dans l'ensemble, ces résultats renforcent l'idée centrale selon laquelle tous les gestes devraient être caractérisés en fonction de trois dimensions largement indépendantes : leur forme, leurs caractéristiques prosodiques et leur signification sémantique et/ou pragmatique, ce qui met au premier plan les trois règles de synchronisation de McNeill.

Les résultats de ces études contribuent également à affiner deux des règles de synchronie. En ce qui concerne la règle de synchronie phonologique, les résultats soulignent des aspects méthodologiques importants que les futures études devraient prendre en considération, à savoir que les taux d'alignement pour les apex se sont avérés beaucoup plus variables que pour ceux des strokes. Il est important de noter que les études menées dans le cadre de cette thèse sont les premières à montrer un effet de frontière gauche, où les gestes dans deux langues typologiquement distinctes ont tendance à marquer la frontière gauche de la phrase prosodique, ce qui montre que l'accentuation nucléaire n'est pas le seul point d'ancrage prosodique. En outre, cette thèse (**Chapitre 3**) est la première à montrer que la production rythmique des gestes est toujours guidée par la durée du phrasé prosodique. En ce qui concerne la règle de la synchronisation pragmatique, cette thèse permet de montrer que la parole et le geste ne font pas que véhiculer le même sens pragmatique. En effet, les études de cette thèse ont révélé que les deux modes peuvent aussi interagir de manière complémentaire pour transmettre une même intention pragmatique.

Dans l'ensemble, les quatre études empiriques de cette thèse contribuent à notre compréhension de l'association prosodique et des fonctions pragmatiques des gestes, mais démontrent également comment ces caractéristiques prosodiques et pragmatiques ne doivent pas être considérées comme dénotant une typologie mutuellement exclusive des gestes. Ces résultats empiriques valident le système M3D pour l'étude du langage multimodal, qui se centre sur l'interaction entre les modes de communication et sur une analyse des gestes basée sur trois dimensions indépendantes et non mutuellement exclusives. En définitive, cette thèse milite en faveur de l’adoption du système M3D pour une approche standardisée et multidisciplinaire de l'étude des gestes.

Table of Contents

	Pàg.
Acknowledgements.....	v
Abstract.....	xii
Resum.....	xvi
Resumen.....	xx
Sommaire.....	xxiv
1. GENERAL INTRODUCTION.....	1
1.1. Why study co-speech gesture?.....	2
1.2. The study of co-speech gesture: Contributions of Kendon and McNeill.....	7
1.2.1. What is a gesture?.....	7
1.2.2. Gesture classification.....	9
1.2.3. The organization of movement (gesture units and gesture phases).....	15
1.2.4. Issues with the current theoretical approaches to gesture classification: A more recent view….....	18
1.3. The temporal relationship between gesture and prosody…….....	26
1.3.1. The Autosegmental-Metrical (AM) system….....	27
1.3.2. Gesture and its temporal association with prosodic structure…….……	37
1.4. Multimodal cues to information structure……...	53
1.4.1. General overview of information structure……..	54
1.4.2. Prosodic marking of information structure……..	60
1.4.3. Gestural marking of information structure……....	64
1.5. General objectives, research questions, and hypotheses……...	71
2. MULTIDIMENSIONAL LABELING OF GESTURE IN COMMUNICATION — THE M3D PROPOSAL……...……	79
2.1. Introduction…….	80
2.1.1. A holistic view of multimodality in language…..	81
2.1.2. A multidimensional approach to gesture labeling…….	85
2.1.3. Main features of the currently available gesture labeling systems…….……..	89
2.1.4. Main goals……	92

2.2. Survey of existing multimodal annotation systems…… 94
2.2.1. Goals of the target multimodal annotation systems……...…….. 95
2.2.2. Coding of multimodal language: speech and speech prosody……..…… 96
2.2.3. Summary of coverage of gesture features across multimodal annotation systems…… 99
2.2.4. Accessibility, explicitness and applicability…… 105
2.3. Main features of the M3D annotation system……. 111
2.3.1. The concept of multimodal labeling…… 111
2.3.2. Multidimensional view of gesture…… 112
2.3.3. Gesture annotation……... 114
2.3.4. M3D accessibility, explicitness and applicability…….. 133
2.3.5. M3D_TED French and English Corpora……. 137
2.3.6. Reliability of key aspects of the M3D annotation……. 143
2.4. Discussion and conclusions…… 150

3. PHRASAL PROSODIC STRUCTURE AS A KEY FACTOR IN THE TEMPORAL EXECUTION OF GESTURE IN FRENCH ACADEMIC DISCOURSES……. 153
3.1. Introduction……. 154
3.1.1. The temporal association between prominence in speech and gesture……... 156
3.1.2. A brief overview of French prosodic structure… 158
3.1.3. Studies on the temporal association between gesture and prosody in French……. 160
3.1.4. Gestural rhythm…… 161
3.1.5. Motivation and research questions……... 163
3.2. Methods……... 166
3.2.1. Materials: The French M3D-TED Corpus……... 166
3.2.2. Data annotation…… 167
3.2.3. Gesture annotation……... 167
3.2.4. Prosodic annotation…….. 173
3.2.5. Gesture-speech alignment criteria…… 174
3.2.6. Statistical analyses……... 175
3.3. Results……. 177
3.3.1. Temporal alignment between manual gesture

strokes and apexes with pitch accented syllables……..... 177
3.3.2. The rhythmic productions of referential and non-referential gestures…… 181
3.3.3. The relationship between RGGs and APs…… 182
3.4. Discussion and conclusions…… 186

4. VISUALIZING PROSODIC STRUCTURE – MANUAL GESTURES AS HIGHLIGHTERS OF PROSODIC HEADS AND EDGES IN ENGLISH ACADEMIC DISCOURSES..... 199
4.1. Introduction..... 200
4.1.1. Gesture types, landmarks, and their association with prosodic prominence..... 201
4.1.2. The role of phrasal prosodic structure in gesture production……..... 208
4.1.3. Motivation and research questions..... 212
4.2. Methods……... 214
4.2.1. Materials: The English M3D-TED Corpus…….. 214
4.2.2. Data annotation…… 215
4.2.3. Gestural annotation…….. 216
4.2.4. Prosodic annotation……. 219
4.2.5. Gesture-speech alignment criteria…… 222
4.2.6. Statistical analyses……... 223
4.3. Results……. 225
4.3.1. Temporal alignment between manual gesture strokes and apexes with pitch accented syllables…… 225
4.3.2. Temporal association between manual gesture strokes and prenuclear and nuclear pitch accentuation……. 228
4.3.3. Gestural attraction towards the prenuclear pitch accents: An effect of relative prominence?..... 231
4.3.4. Gestural attraction towards prenuclear pitch accents: A phrase-initial edge strengthening effect?..... 234
4.4. Discussion and conclusions…… 238

5. THE MULTIMODAL MARKING OF INFORMATION STATUS OF REFERENTS IN ENGLISH ACADEMIC DISCOURSES……... 249

5.1. Introduction……. 250
5.1.1. Information structure and the information status of discourse referents……... 251
5.1.2. The prosodic marking of ISR..... 255
5.1.3. The gestural marking of ISR…… 259
5.1.4. Motivation and research questions……... 262
5.2. Methods……... 264
5.2.1. Materials: The English M3D-TED Corpus…….. 264
5.2.2. Data annotation…… 265
5.2.3. Gestural annotation…….. 265
5.2.4. Prosodic annotation…….. 267
5.2.5. ISR annotation…….. 271
5.2.6. Assessing multimodal cues to ISR……... 272
5.2.7. Statistical analyses……... 274
5.3. Results……. 277
5.3.1. Gesture and prosody as joint cues for ISR……... 277
5.3.2. Prosodic cues: Disentangling pitch accent type and relative degree of prominence……... 280
5.3.3. Gestural cues: Differences by gesture type and their role in prenuclear positions…….. 282
5.4. Discussion and conclusions…… 284

6. GENERAL DISCUSSION AND CONCLUSIONS.. 295
6.1. Summary of findings……... 296
6.2. The value of M3D……... 302
6.3. Refining the phonological synchrony rule…….. 306
6.3.1. The association of strokes with pitch accented syllables: High levels of synchronization across gesture types……. 307
6.3.2. The role of the apex in the temporal relationship between gesture and pitch accentuation……... 309
6.3.3. The effects of phrasal prosodic structure: The important role of phrasal position……... 312
6.3.4. The effects of phrasal prosodic structure: Phrasal duration predicts distance between RGG apexes……..……... 316
6.4. Refining the pragmatic synchrony rule……... 319
6.4.1. Multimodal cues to ISR……... 319

6.5. Future work for M3D…….. 323
6.6. Final conclusions……. 325

REFERENCES…….. 327

1

CHAPTER 1: GENERAL INTRODUCTION

1.1. Why study co-speech gesture?

When humans communicate with each other, they make use of a wide variety of strategies to convey the message so that our interlocutor can accurately and easily understand the meaning they wish to convey. In oral communication, they rely not only on a string of sounds which represent words, structured into a codified morphosyntactic grammar, but we also superimpose meaning through intonational and rhythmic patterns, and can use our body to help vehicle this message for the correct interpretation of the utterance. For a large part of its history, the study of linguistics has mostly focused on speech or text, specifically formalizing phonological, morphological and syntactic structures. Meanwhile, gestures, facial expressions, and bodily movements have been considered non-linguistic aspects of communication, and have been studied as a "separate" domain from language. The study of linguistics has since come to include broader aspects such as social interaction and embodiment (see Mondada, 2016; Perniss, 2018 for a review). Before going into detail about theoretical approaches to the study of gesture (that is, *how* researchers study gesture), it is particularly important to understand exactly *why* gesture is worthy of study. Thus, the first subsection of this thesis aims to expose a (very small) peek at some of the main findings across various disciplines that have led to the incorporation of gesture as a field of linguistic study, highlighting that language is multimodal.

First, since the early 90's, gesture has been argued to be a key component of language. McNeill's work has largely been framed

through a psycholinguistic perspective, where he has given evidence that gesture is "a window onto thought." His seminal 1992 book largely focused on how gesture and speech are integrated into a single system, and how gestures reveal the thought processes that are going through a speaker's mind when communicating. One of his main arguments has been dubbed the "three synchrony rules" which describe how gesture and speech are temporally executed together (*phonological synchrony rule*), semantically coherent (*semantic synchrony rule*), and pragmatically coherent (*pragmatic synchrony rule*).

The phonological synchrony rule developed from the earlier observations by Kendon (1980) that the gesture stroke (i.e., the obligatory movement phase of a gesture which bears meaning) "precedes or ends at, but does not follow, the phonological peak syllable of speech" (McNeill, 1992, p. 26). When this rule is threatened, speakers have been shown to temporarily stop gesturing so as to maintain phonological synchrony, as well as in cases of disfluency (e.g., Graziano & Gullberg, 2018). According to the semantic synchrony rule, speech and gesture should be reflecting the same idea since they are conceptualized together. However, the synchrony behind the semantic meaning portrayed in speech and gesture is not always straight-forward. Rather, the relationship can be seen as falling on a continuum where at one end, gesture represents exactly the semantic content in speech (such as saying the word "driving" and holding up one's hands as if handling a steering wheel; see Bergmann et al., 2011). These gestures are

referred to as “redundant” gestures. On the other side of the continuum, gesture can be “complementary” or "non-redundant" in that it represents semantic content which is absent in speech (for example, if the same “driving” gesture is produced, yet the speaker says “I went to the supermarket”). In such contexts, the speaker is adding additional information about *how* they arrived at the supermarket (by car and not by bike). Thus, gesture and speech “jointly express the same core meaning and highlight different aspects of it” (McNeill, 2000, p. 7). Finally, McNeill describes the pragmatic synchrony rule explaining how gesture and speech share the same pragmatic function. For example, he describes how the utterance “… it was a Sylvester and Tweety cartoon”, was produced by a speaker who also gestures depicting a bounded object. The pragmatic synchrony between the two modes indicate that both are functioning to introduce the genre of the upcoming discourse. In addition to the synchrony rules, McNeill describes how gestures develop together with speech, and both speech and gesture break down together in aphasia.

Furthermore, more recent studies have shown how gestures are integrated and processed with speech together at the neural level. Studies using neuroimaging techniques such as Electroencephalography (EEG) and functional Magnetic Resonance imaging (fMRI) have shown that gesture and speech are processed similarly and in a holistic fashion in terms of semantic meaning (e.g., Kelly et al., 2004; Özyürek et al., 2007) and prosodic meaning (e.g., Biau et al., 2016; Dimitrova et al., 2016; Hubbard et al.,

2009). Furthermore, processing is sensitive to the temporal association between gesture and co-expressive speech (Habets et al., 2011) and takes place in similar regions of the brain (e.g., Andric et al., 2013; Wolf et al., 2017). Additionally, gesture has been shown to even speed up speech processing (Skipper, 2014; Weisberg et al., 2017).

Second, gesture has also become key in developmental research. Indeed, children make use of gesture for communication before they learn to speak. Specifically, children begin producing their first deictic (i.e., pointing) gestures around 9 to 12 months of age (see Rohlfing et al., 2017, for a review). As children grow and acquire language, not only do both speech and gesture develop together (e.g., McNeill, 1992; Özçalışkan & Goldin-Meadow, 2005), but gesture actually is a forerunner, often showing development before speech. Moreover, it signals upcoming changes, predicting speech and cognitive development. For example, the production of gesture with single-word utterances predicts the onset of two-word utterances (Butcher & Goldin-Meadow, 2000; Capirci et al., 1996; Goldin-Meadow & Butcher, 2003). Later in development, gesture can predict other linguistic skills, such as lexical and grammatical development (e.g., Igualada et al., 2015), narrative ability (Demir et al., 2015; Vilà-Giménez et al., 2021) and pragmatic competence (Hübscher & Prieto, 2019; Pronina et al., 2021). This predictive effect also applies to domains outside of language like the expression of Piagetian conservation of quantity (Church & Goldin-Meadow, 1986) and mathematical equivalence (Alibali & Goldin-

Meadow, 1993), where gesture can signal that children are ready to learn abstract concepts. In addition to first language development, gesture has also proven to have an important role in the acquisition of a second language, where they can help with novel word learning (e.g., Tellier, 2008; Kushch et al., 2018; see Macedonia, 2014 for a review), make corrective feedback from instructors more efficient (Nakatsukasa, 2016), and boost phonological learning (i.e., pronunciation) both at the segmental (e.g., Li et al., 2020) and suprasegmental levels (Llanes-Coromina et al., 2018; Baills et al., 2019; see also Baills et al., 2022 for a review).

Psycho- and neurolinguistic studies have shown how gestures boost cognition. For example, gestures boost problem-solving in mathematical tasks (Broaders et al., 2007; Cook et al., 2008; Goldin-Meadow et al., 2009; Novack et al., 2014) as well as in spatial thinking tasks (Alibali & Kita, 2010; Alibali et al., 2011). In terms of memory, not only is seeing gestures beneficial (Cohen & Otterbein, 1992; Feyereisen, 1998; Austin & Sweller, 2014; Macoun & Sweller, 2016; Llanes-Coromina et al., 2018), but speakers who produce gestures at encoding are better at later recall (e.g., Cook et al., 2012; Wagner et al., 2004; Morett, 2014). Gestures also boost comprehension of speech (Driskell & Radtke, 2003; McNeil et al., 2000) as well as narratives (Dargue & Sweller, 2019; Macoun & Sweller, 2016; Llanes-Coromina et al., 2018; see Vilà-Giménez & Prieto, 2021 for a review).
The current subsection is in no sense an exhaustive review of the studies that overwhelmingly show that gesture is indeed a key part

of language and interacts with speech in development, learning, and cognition. The field of gesture studies has developed to be an important subfield of linguistics that is largely interdisciplinary (as evidenced by the range of studies described in this subsection) and espouses a multimodal view of language. In the present Ph.D. thesis, we support a more comprehensive view of "*multimodal language*" that refers to the use of multiple modes of communication as meaning-making strategies which include aspects such as speech prosody, gaze, manual gesture, facial expressions, body postures, etc. (see, e.g., Goodwin, 2000; Mondada, 2016; Perniss, 2018). Importantly, the foundations of these previous studies are based on different theoretical approaches regarding how to identify, classify, and describe gestures. **Subsection 1.2** thus aims to describe *what is* gesture and the various approaches used in studying gestural phenomena.

1.2. The study of co-speech gesture: Contributions of Kendon and McNeill

1.2.1. What is a gesture?

Though the study of gesture has been around for centuries (for a review of the history of gesture studies, see Kendon, 2004; 2017), it was only in the past century that researchers really took an interest in co-speech gesture as a central component to human language. Specifically, *co-speech gesture* refers to "a visible action of any body part, when it is used as an utterance, or as part of an utterance" (Kendon, 2004, p. 7). By this definition, co-speech gesture does not

include other paralinguistic body movements which do not convey a communicative intention, such as touching one's hair, shifting the body position for comfort, scratching, etc. (the literature generally uses the term *adaptors* to refer to such movements, e.g., Butterworth et al., 1981; McNeill, 1992).

In the past decades, the works by Adam Kendon and David McNeill represent some of the most foundational and widely accepted works that are used today for gesture studies. Kendon was among the first to begin to distinguish and categorize gestures (1980, 1982), which in turn inspired McNeill (1992, 2000) to further develop the *Kendon Continuum*. This continuum largely identifies four types of communicative movement: gesticulation, pantomime, emblem, and sign language. They are arranged in such a way that "As we move from left to right: (1) the obligatory presence of speech declines, (2) the presence of language properties increases, and (3) idiosyncratic gestures are replaced by socially regulated signs." (McNeill, 1992, p. 37). *Gesticulation* refers to more spontaneous communicative movements. For example, if a speaker utters "He grabs a big oak tree and bends it way back" while holding one arm upright and then pulling it back with the other arm to represent bending a tree backward towards the ground (example taken from McNeill, 2000). At this end of the continuum, gestural meaning is interpreted through the context of concurrent speech and is determined in a top-down (or global) fashion, that is, "the meanings of the 'parts' are determined by the meaning of the whole" (McNeill, 2000, p. 5). *Pantomimes* are gestures which are not co-produced with speech. For example, if a person replies to the question "what's a vortex?"

by simply twirling their index finger in a circle. *Emblems* are highly conventionalized and culture-specific (e.g., the "peace sign," or "thumbs up" to signal positive confirmation). Emblems may or may not be produced with speech. The features of these emblems contrast with the *signs* used in sign languages, which are both highly conventionalized and contain full linguistic properties. Their gestural meaning is determined in a bottom-up fashion, in that individual morphemes (constrained by, e.g., morphosyntactic and phonological properties) come together to create meaning. In other words, the meaning of the parts determines the meaning of the whole.

To sum up, the field of gesture studies is largely based on Kendon's initial observations regarding the relationship between speech and gesture, which was in turn further developed by McNeill's work. Specifically, McNeill (1992, 2000) builds upon Kendon's work by developing the Kendon Continuum as well as a classification system that is widely used in the field today, where gestures generally refer to the left end of the Kendon continuum (i.e., gesticulation, pantomime, emblem). For the remainder of this thesis, the term (co-speech) gesture will refer to these left-most categories in the Kendon Continuum (unless otherwise specifically mentioned).

1.2.2. Gesture classification

Further in line with McNeill's psycholinguistic perspective of showing how gesture is a window onto thought (see **subsection**

1.1), he developed a classification system that is among the most widely used in the field today. It largely distinguished gestures into four main types: iconic, metaphoric, deictic, and beat gestures. Iconic gestures are those which pictorially represent a concrete action or object, such as the "driving" gesture described above, bearing a close relationship with the semantic meaning in speech which it accompanies. Metaphoric gestures are similar in that they pictorially represent semantic content, but they represent abstract concepts or ideas. For example, uttering "the trip will take three days" while producing a gesture with a flat hand moving to the right, representing time (the abstract concept of three days) on a physical axis (see also Cienki & Müller, 2008). Deictic gestures refer to familiar pointing gestures. These may be used in a concrete manner (to locate something present in the immediate environment) or an abstract manner. Abstract pointing is particularly important in narrative speech, where speakers may situate discourse referents or entities (which are not immediately present in the environment) in a locus in space which serves as a spatial reference point for future mentions (see, e.g., Gullberg, 1998). For example, if a speaker introduces two discourse referents, saying "There was a dog and a cat.", the speaker may point towards the left to indicate the dog, and to the right to indicate the cat. Further in the discourse, when the speaker mentions the cat, they may reproduce a pointing gesture to the space that was previously assigned to the cat, aiding in referent tracking for the listener.

The final gesture category described by McNeill is the beat gesture. These gestures do not represent semantic content nor do they

spatially refer to entities via pointing. Instead, these gestures have largely been classified based on their form. Specifically, McNeill (1992) describes:

> The hand moves along with the rhythmical pulsation of speech [...] Unlike iconics and metaphorics, beats tend to have the same form regardless of the content [...] The typical beat is a simple flick of the hand or fingers up and down, or back and forth; the movement is short and quick and the space may be the periphery of the gesture space (the lap, an armrest of the chair, etc.). (p.15)

He goes on to say that these gestures can be identified through their simple biphasic nature, claiming that they often contain only two movement phases (up/down, in/out), whereas iconics and metaphorics typically contain three (preparation - stroke - recovery). These characteristics form the basis of the "beat filter" (p. 81), a series of questions that annotators may use to assess whether a gesture is imagistic (i.e., iconic or metaphoric) or not (i.e., a beat). For each question, a yes answer receives one point, and a no answer receives zero points:

1. Does the gesture have other than two movement phases (i.e., either one phase, or three phases, or more)?
2. How many times does wrist or finger movement or tensed stasis appear in any movement phase not ending in a rest position (add this number to the score).

3. If the first movement is in a non-center part of space, is any other movement performed in center space?
4. If there are exactly two movement phases, is the space of the first phase different from the space of the second?

If a score of 0 is returned, the gesture is most likely a beat, while a score of 5 or 6 is more likely indicative of an iconic or metaphoric gesture.

In addition to being defined by their form, beat gestures have also been described in terms of their discourse-pragmatic value. While beat gestures do not convey semantic meaning, they index the words or phrases they accompany for their relevant discourse-pragmatic content. They have specifically been ascribed multiple functions in narratives by McNeill (1992), such as marking the introduction of new characters, summarizing action or introducing new themes. Such functions allow for the structuring of the discourse and for events on the meta-narrative level to be directly inserted into the narrative itself (indicating departures from the narrated chain of events).

This classification is potentially the most widely used system in the field of gesture studies (though it is not the only one). This conceptualization of "gesture types" has led to an interpretation of mutually-exclusive categories (i.e., a gesture should fit into one and only one of the four potential categories). In 2006, McNeill clarified his position, claiming that these types should rather be seen as "dimensions" of meaning, where gestures may portray different

levels of “iconicity”, “metaphoricity”, “deixis” or “temporal marking”. Specifically, he says:

> The essential clue that these are dimensions and not categories is that we often find iconicity, metaphoricity, deixis and other features mixing in the same gesture. Beats often combine with pointing, and many iconic gestures are also deictic. We cannot put them into a hierarchy without saying which categories are dominant, and in general this is impossible. A practical result of dimensionalizing is improvement in gesture coding, because it is no longer necessary to make forced decisions to fit each gesture occurrence into a single box. (McNeill, 2006, p. 60)

Figure 1.1 gives an example of a manual gesture that occurs within a series of five subsequent manual gestures showing degrees of iconicity, metaphoricity, and temporal marking. The speaker says “Finally, we could make the particles migrate to **over** the **poles**, so we could arrange the climate engineering so it **really focused** on the **poles**” (**bold** indicates words co-occurring with a gesture). The gesturing shows iconicity, as it iconically represents poles sticking out at the ends of the Earth. The gesturing is also metaphoric in that it is representing an abstract concept, as the poles of the earth are not actual poles but rather refer to specific geographic/magnetic points on Earth. Finally, the speaker produces the same gesture multiple times along with prominent syllables in speech, seeming to loosely mark speech rhythm.

Figure 1.1: An example of a single gesture taken from a series of gestures which contain iconicity, metaphoricity, and temporal marking taken from Keith (2007) at 08:34.

The McNeillian approach is but one of many potential manners of classifying gesture. In fact, McNeill largely based his classification on systems that were developed before (e.g., Efron, 1941; Ekman & Friesen, 1969; Freedman & Hoffman, 1967; See Kendon, 2004 for a review). Another key method of typologizing gesture was that described by Kendon (2004). Not a "classification system" per se, he describes how different gestures work to contribute meaning to an utterance. He first described gestures that have a referential function – that is, they contribute meaning by representation or pointing. Gestures that do not fulfill a referential function are said to fulfill a pragmatic function. Interestingly, in his 2004 book, Kendon concedes that perhaps the term "pragmatic" to denote any gestures which do not convey a referential meaning may not be ideal, but says that "no other terms seems available." (p. 159). In Kendon

(2017), the author develops his proposal about pragmatic gestures and describes the four pragmatic functions of gesture, describing an operational function (affirmation/negation), a modal function (to provide an interpretive frame for a stretch of speech), a performative function (manifesting speech acts), and a parsing function (marking discourse structure).

Recent views have called for an approach to gesture classification that recognizes that semantic, pragmatic, and prosodic properties of gestures should be accounted for in an independent manner (e.g., Prieto et al., 2018; Prieto & Shattuck-Hufnagel, 2019). **Subsection 1.2.4**. will explore the motivation behind such a view. The upcoming **subsection 1.2.3.** will describe commonly-held approaches to describing gestural movements.

1.2.3. The organization of movement (gesture units and gesture phases)

Manual co-speech gestural movements can be segmented into a hierarchical fashion (Kendon, 1980; Kita et al., 1998). Specifically, the smallest and only obligatory gesture phase is the *stroke*. This phase is generally characterized as a peak of "effort" or "accented movement," and is generally considered the meaningful part of the gesture (McNeill, 1992). Strokes may be preceded by a *preparation* (the movement of the hands into position to execute the stroke), or followed by a *recovery* (sometimes called *retraction*, referring to a return of the hands to rest). Additionally, *holds* (or moments of minimal movement) may occur before or after the stroke. The

combination of a single stroke and any other associated gesture phases (e.g., preparation, hold, etc.) is considered a *Gesticular Phrase* (henceforth, G-Phrase). G-Phrases can then concatenate one after another, in which immediately subsequent G-Phrases can be grouped into a single, larger unit, termed the *Gesticular Unit* (G-Unit). The G-Unit thus coincides with the moment from which the hands leave a rest position until their return, the span over which a speaker may produce a single or multiple concatenating gestures (see **Figure 1.2**).

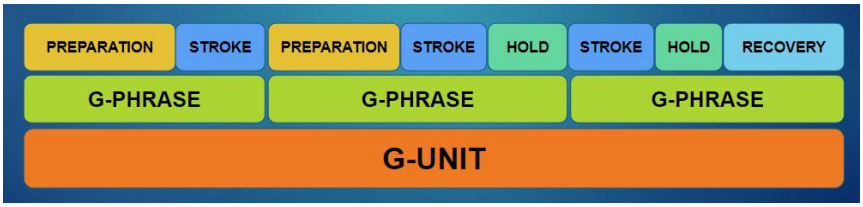

Figure 1.2: Schematic representation of gesture phases (preparation, stroke, hold, recovery) and their grouping into larger G-Phrases and G-Units.

Some researchers have also identified another gestural landmark within the stroke. The *apex* (Loehr, 2004; 2007; also termed *hits* by others, e.g., Yasinnik et al., 2004) refers to a single moment in time corresponding to the maximum "peak" of the stroke or "kinetic goal" of the stroke. If a stroke is *unidirectional* (i.e., the stroke only contains a single movement in one direction, such as a pointing gesture), the endpoint of the stroke is considered the goal of the stroke, and thus is marked as the apex. *Bi-directional* strokes (e.g., down/up or out/in movement), the point where the direction changes is marked as the apex. When strokes are made up of multiple movement directions (i.e., *multi-directional*), multiple

apexes are identified. **Figure 1.3** shows an example of a gesture executed within a gesture unit. The upper panel shows how the speaker begins the gesture from a hold position, moves his hands upward (preparation), then quickly downward (stroke) before stopping in a hold before a subsequent gesture. The lower panel shows frame-by-frame screenshots of the stroke, and identifies the apex of the stroke (the screen where the left hand suddenly clears up, indicating zero velocity).

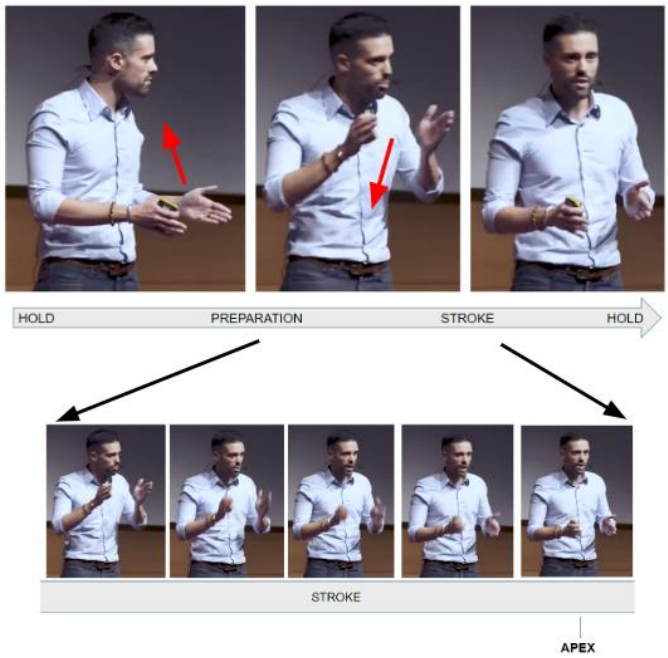

Figure 1.3: Still images of a (non-referential) gesture executed in the French M3D-TED corpus, by speaker JP (TEDx Talks, 2018) at 02:58. **Upper panel**: the various gesture phases involved in the execution of the gesture. **Lower panel**: frame-by-frame images of the stroke, where the final frame indicates the apex.

1.2.4. Issues with the current theoretical approaches to gesture classification: A more recent view

In the present section, we would like to highlight three aspects of McNeill's (1992) theoretical view regarding gesture that currently is in need of some refinement by taking on board more empirical investigation, namely the phonological synchrony rule, the pragmatic synchrony rule, and the gesture classification proposal.

1.2.4.1. The phonological synchrony rule

As previously mentioned in **subsection 1.1.**, the phonological synchrony rule holds that "the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech." (McNeill, 1992, p. 26). This rule was largely based on Kendon (1980) who describes how gestures tended to associate with the nuclear stress of the Tone Unit (following the prosodic model by Crystal & Davy, 1969). More precisely, the results of Kendon's (1980) analysis showed that of 22 Tone Units which aligned with a gesture, 15 strokes were completed before or at the onset of the nuclear syllable, six strokes were completed by the end of the nuclear syllable (representing overlap between the nuclear syllable and stroke) and finally only one stroke continued beyond the nuclear syllable. Thus, these initial observations indicate a rather loose or imprecise relationship with nuclear pitch accentuation (for example, of the majority of strokes that occurred before the nuclear syllable, did they also co-occur with other prominences in speech?). Importantly, when taking on this generalization, McNeill's

phonological synchrony does not explicitly describe what the "phonological peak syllable" refers to, or exactly what is the domain of prosodic phrasing in which this relationship exists (see **subsection 1.3.1.** of this chapter). It any case, it seems to make a clear connection between gesture production and prosodic prominence (i.e., the salience of certain syllables over others).

All in all, McNeill's (1992) phonological synchrony rule is not anchored in a complete prosodic model, which makes it rather difficult to interpret, as well as to make predictions from a language typology perspective. Unfortunately the lack of precision of the model, together with the abovementioned initial results and others, have led to a tendency to focus on the role of prosodic prominence alone as the main attractor for gesture production, and more specifically to the belief that it is only the nuclear syllable (e.g., the most structurally prominent syllable within the prosodic phrase) that acts as an anchor unit for gesture production (e.g., Ebert et al., 2011; see **subsections 1.3.1. and 1.3.2.** of this chapter for further discussion). As we will see, the lack of precision in the phonological synchrony rule is an issue that more recent research has started to address and that needs further investigation (see a review of the literature in **subsection 1.3.2.**).

1.2.4.2. The pragmatic functions of gestures

The pragmatic synchrony rule holds that speech and gesture should be pragmatically coherent. While McNeill (1992) does not dedicate much of his work to describing the pragmatic relationship between

speech and gesture much further (apart from discussing some aspects of discourse structure and Communicative Dynamism, see **subsection 1.4.3.**), numerous studies have assessed the pragmatic functions of gestures (see **Chapter 2** of this thesis for an overview). However, two main criticisms can be issued in this regard, namely (a) these studies have not developed a systematic approach to describing the pragmatic functions of gesture and thus there is a need to integrate current views in the pragmatics field that can help elucidate the multimodal expression of pragmatic meaning; (b) while the phonological and semantic synchrony rules have since been further refined (e.g., further clarifications on how gestures associate with prominence making use of empirical laboratory studies or kinematic measures; the wide array of studies showing how gestures may convey semantic meaning in a variety of ways), the pragmatic synchrony rule has seen much less development and there is a need for further experimental studies that allow us to assess how gesture and speech interact to convey pragmatic meaning.

1.2.4.3. McNeill's gesture classification in the face of the three synchrony rules

As previously mentioned, the most widely-used theoretical approach to classifying gesture is that of McNeill (1992), which distinguishes iconic, metaphoric, deictic, and beat gestures. The division of gestures into these four categories was done by taking separate criteria, where the former three have a clear referent in speech, while the latter "tend[s] to have the same form regardless of

content", being "flicks of the hand" which "seem[s] to be beating musical time" (McNeill, 1992, p. 15).

A first issue is in terms of gestural form, as recent corpus-based studies have shown that beat gestures may indeed have complex phasing and may have a variety of hand forms and trajectories (e.g., Shattuck-Hufnagel et al., 2016). A second issue is that while some gestures are classified, categorized, and named for their semantic relationship with speech, the name "beat gesture" (and McNeill's original definition, see above) implies that only this gesture type has a close relationship with prosody and rhythm. However, recent studies have shown that beat gestures are not always coupled with prominence in speech, and they associate with prominence at similar rates as referential gestures (e.g., Shattuck-Hufnagel & Ren, 2019; Rohrer et al., 2019).

A third issue relates to their semantic properties in discourse. Referential gestures are said to be "meaningful" in that they clearly convey semantic content in speech, yet beat gestures have been described as "meaningless" in the literature due to their lack of *semantic* meaning (see, e.g., Abner et al., 2015; Weisberg et al, 2017; but for contrasting views on their semantic contribution, see Yap & Casasanto, 2018, for how beat gestures may encode spatial semantics). Importantly, on the other hand, beat gestures have already been linked to various pragmatic functions and meanings (such as introducing or summarizing information, see McNeill's description of beat gestures above; see also Loehr, 2012). Some researchers have also claimed that they may also be considered as

interactive conversational gestures, as they contribute to the "nature of dialogue itself, rather than with the specific topic of discourse" (Bavelas et al., 1992, p. 476).

Crucially, not only do beat gestures fulfill pragmatic functions, but referential gestures do as well. For example, many of the examples Kendon describes in this (2017) work on the pragmatic functions of gestures are themselves referential in nature (e.g., the "air quote" gesture). Moreover, many studies can be found regarding the pragmatic functions of the "Palm Up Open Hand" gesture (e.g., Ferré, 2011). Importantly, these gestures are realized as an open hand with the palm up and placed in front of the speaker as if to offer something, and are often interpreted as metaphoric as they "treat the abstract objects of discourse—propositions, ideas, questions, answers—like the physical objects of everyday life in that they can be held up, offered, requested, exchanged, and so on." (Cooperrider et al., 2018, p. 5; see also Cienki & Müller, 2008). Furthermore, gesture is increasingly seen as polyfunctional, in that a single gesture may accomplish multiple pragmatic functions at once (see, e.g., Lopez-Ozieblo, 2020).

Following one of McNeill's central points regarding the cognitive links between brain and gesture (specifically, that all gestures regardless of type show a semantic, pragmatic, and phonological synchrony with speech), it is not accurate to highlight the exclusive prosodic and pragmatic role of beat gestures and subsequently classify all other gestures in terms of their semantic properties. The name "beat" gestures is thus misleading and unfortunately

contributes to the belief that there is one particular type of gesture that associates more with prosodic prominence and plays key discourse-pragmatic functions in speech. As we mentioned, some work has shown that beat gestures are no more related to prosody than other gesture types, and other gesture types also contribute pragmatic meaning in discourse. In the next subsection we will describe a more recent view on how to typologize gesture that is based on the assessment of the independent semantic, pragmatic, and prosodic properties of gesture.

1.2.4.4. A recent multidimensional view of gesture

In 2006, McNeill discouraged a categorical approach to his classification, and espoused a dimensionalized approach, where gestures can be seen as consisting of iconic, metaphoric, deictic, and "temporal highlighting" (beat) dimensions (or any mixtures thereof). However, this dimensionalized approach is still problematic, as it continues to conflate various aspects of semantic meaning with association with prosodic prominence, and does not specify how to assess gestures which do not convey semantic meaning nor associate with prosodic prominence. Furthermore, it does not take any pragmatic assessment into account. Crucially, in our view, a "temporal highlighting dimension" should be seen as *independent* of different semantic dimensions of referentiality (i.e., semantic contributions of iconicity, metaphoricity, and deixis) or pragmatic function. Following up on McNeill's (2006) proposal, recent views have begun calling for a more comprehensive dimensionalized approach to studying gesture (e.g., Prieto et al.,

2018; Shattuck-Hufnagel & Prieto, 2019). Such a view puts McNeill's three synchrony rules at the forefront by recognizing the semantic, pragmatic, and prosodic properties of gestures, and applies them to every gesture in an independent fashion. In a practical sense, this proposal suggests that, researchers should assess the semantic coherence with speech (whether they are *referential* in that they refer to semantic content – via iconicity, metaphoricity, and deixis, such as in **Figure 1.3** – or are *non-referential* in that they do not, such as in **Figure 1.2**), the pragmatic coherence with speech (whether they contribute to pragmatic meaning – or not), and their prosodic association with speech (whether they are produced in synchrony with prosodic structure – or not).

The current thesis will thus be anchored in such a multidimensional approach to the study of gesture, and aims to build upon these ideas by offering the novel MultiModal MultiDimensional (M3D) labeling system to implement this recent approach to the studies of gesture for the development of multimodal corpora (**Chapter 2** in this thesis). By offering openly accessible, easy to follow guidelines for the annotation of co-speech gestures, researchers not only in the field of gesture studies, but also in other disciplines, will be able to grasp much more precise details about multimodal human communication, specifically by accounting for the multiple dimensions of gesture production, and how they are integrated with the multiple channels through which speakers convey information.

With respect to gesture classification, the current thesis will make use of a more general distinction between referential and non-referential gestures. This proposal follows work by Prieto et al. (2017), Shattuck-Hufnagel & Prieto (2019) and Shattuck-Hufnagel & Ren (2018) that indicated a basic divide between gestures that convey semantic meaning (e.g., representational and deictic gestures) and those that not convey semantic meaning in speech (see **Chapter 2** for more details). Using this basic distinction between gesture types has the advantage of avoiding the strict definition of beat gestures put forth by McNeill (1992) and including all manual gestures of various forms. Despite this, for the sake of clarity and transparency, the present review of the literature exposed in the introductory sections of this thesis (**Chapter 1**; introductory sections of each subsequent chapter) will use the terminology originally employed by the authors of the referenced studies when discussing gesture types.

In sum, following up on the present state of the art, **Chapter 2** will introduce the MultiModal MultiDimensional (M3D) approach to gesture labeling. The subsequent chapters will demonstrate how applying M3D to two corpora of English and French TED Talks can further refine our knowledge on how gesture is integrated with prosodic structure (**Chapters 3** and **4**), and how prominence in speech and gesture work together to mark information structure (**Chapter 5**). In order to set the stage for these studies, the following subsections will offer a brief introduction to the temporal relationship between gesture and prosody (**subsection 1.3.2.**) and

the (multimodal) marking of information structure (**subsection 1.4.**).

1.3 The temporal relationship between gesture and prosody

The previous section has already introduced the phonological synchrony rule, which holds that gesture strokes occur just before or at the onset of the "phonological peak syllable of speech." As we also mentioned one of the issues is that McNeill's (1992) description is not anchored in a clear prosodic model, which makes it rather difficult to interpret. Thus, the current subsection will begin with a brief description of the Autosegmental-Metrical (AM) theory of intonational phonology and its application in French and English, two typologically different languages from a prosodic point of view that will be studied in the current thesis. Crucially, the fact that French and English differ in their metrical and intonational structure offers an opportunity to better understand the temporal relationship between gesture and prosody that takes into account a phrasal prosodic structure (and not only prosodic prominence), with the goal to further assess and refine the claims of the phonological synchrony rule. The final subsection will then describe the studies that have investigated the relationship between gesture and prosody.

1.3.1. The Autosegmental-Metrical (AM) system

Prosody has been generally described as the "music of language" and englobes such features as prominence, prosodic phrasing, intonation, and rhythm. In the AM approach (Pierrehumbert, 1980; Gussenhoven, 2004; See also Ladd 2008; Arvaniti 2022 for reviews), intonation refers to the meaningful modulation of pitch (acoustically measured as fundamental frequency, *f0*) to linguistic structuring and pragmatic means. Importantly, the AM approach has led to the development of several language-specific ToBI (Tones and Breaks Indices) annotation systems, where annotators make use of a set of conventional symbols for the annotation of intonation (tones) and phrasing (breaks) (see, Jun 2005; 2014). Specifically, the intonation contour is realized as a sequence of low (L) or high (H) tones which are independent of the segmental string (i.e., they are *autosegments*). Tones associate with structural positions in the metrical representation of an utterance, particularly with constituent heads (i.e., metrically strong syllables) and phrasal boundaries. When tones associate with metrically strong syllables, they are said to be *pitch accents*, and are identified in ToBI annotation with a star, and may be composed of single or complex bitonal movements (e.g., L*, H*, L+H*). In other words, pitch accents refer to intonationally-cued phrase-level prominence. Tones that associate with phrasal boundaries are considered *edge tones*.

1.3.1.1. Prosodic phrasing in French and English

The AM description of French intonational phonology (Jun & Fougeron, 2000, 2002; Delais-Roussarie et al., 2015) describes three levels of prosodic phrasing. The smallest level is called the Accentual Phrase (AP), which is made up of at least one content word and the grammatical words which it governs. APs are designated in the French ToBI system (F_ToBI) as a break index 2. An intermediate level of phrasing has been termed the intermediate phrase (henceforth *ip*) and generally refers to larger prosodic phrases whose occurrence is largely influenced by morphosyntactic structure. Specifically, long branching subject or object NP containing two or more APs, syntactic elements in peripheral positions (e.g., clefted XPs), and non-final elements of an enumeration are often realized as ips. This level of prosodic phrasing also shows relatively larger degrees of phrase-final lengthening compared to APs, and are annotated in the F_ToBI system with a break index 3 and a phrasal accent denoted with a hyphen (e.g., H-; L-). Finally, the largest level of prosodic phrasing is the Intonational Phrase (IP). IPs show a stronger degree of phrase-final lengthening and are often followed by a pause. In speech containing sequences of clauses, each clause is generally realized as an independent IP. IPs are annotated with a break index 4 in the F_ToBI system and a boundary tone denoted with a percent sign (e.g., H%; L%).

In English, two levels of prosodic phrasing can be distinguished: the ip and the IP. Similarly, the two levels of prosodic phrasing are

distinguished by a number of features, such as the degree of juncture (i.e., pause) between subsequent phrases, the degree of lengthening on the final syllable, etc. Additionally, they are annotated similarly in the Mainstream American English ToBI model (MAE-ToBI). Furthermore, these levels of prosodic phrasing are said to be hierarchical, so that each IP must contain at least one ip, and each phrase must contain at least one pitch accent.

1.3.1.2. Domains of pitch accentuation in French and English

As previously mentioned, the association of pitch accentuation is largely determined by metrical structure. According to principles laid out in *metrical phonology* (Liberman, 1975; Liberman & Prince, 1977; Selkirk, 1980, among others), syllables are hierarchically organized into alternating strong and weak prominences, which can be represented on a metrical tree or grid. Representation on a metrical tree involves a binary branching structure at each level of the prosodic hierarchy. The lowest level (relevant for the languages of study in this thesis) corresponds to the syllables. Syllables can then be grouped into feet, which contain at least one strong syllable and any associated weak syllables within the boundary of the lexical word. Above the foot is the prosodic word, which is the domain of primary stress.

Di Cristo (2000; 2016; see also Delais-Roussarie & Di Cristo, 2021 for a review) describes two metrical constraints specific to French: The *Principle of Bipolarization* and the *Principle of Right-dominance*. The Principle of Bipolarization holds that the

underlying metrical structure of content words have prominence at the left and right edge, and the Principle of Right-dominance holds that the right edge is the more prominent. **Figure 1.4** shows a simple grid for the underlying metrical representation of the word "félicité" (bliss).

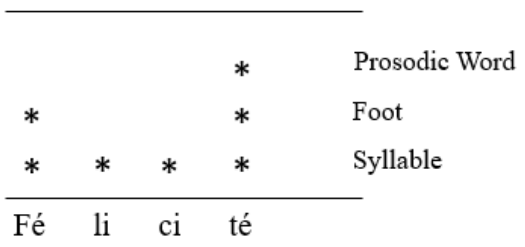

Figure 1.4: Underlying metrical structure for the French word "Félicité" (bliss), adapted from Di Cristo (2000, p. 36)

The metrical constraints described above for French thus act as a template where pitch accents are assigned. Importantly, stress at the lexical level is not cumulative in French, and thus pitch accentuation is assigned at the level of the AP. Each AP (which consists of one or more prosodic words) obligatorily contains a pitch accent on the final non-schwa syllable of a content word (henceforth, FA for final accent). Moreover, the AP may also contain an optional initial accent (IA, coded as Hi in the French ToBI system) on one of the first syllables, marking the left edge of content words (Pasdeloup, 1990; Di Cristo, 1998) or sometimes occurring with grammatical words early in the AP to avoid metrical lapses (Delais-Roussarie, 1996; Jun & Fougeron, 2000). The precise phonological status of the IA is still an area of some debate, where

some researchers consider it a left-edge phrasal accent (e.g., Jun & Fougeron, 2000), a full pitch accent (e.g., Post, 2000), or a "hybrid" accent taking on characteristics of either depending on context (Grice, 2001; Portes et al., 2012). In any case, these two accents have been shown to play a demarcative role, signaling prosodic boundaries at the level of the AP (e.g., Astésano et al., 2007). In addition to having a demarcative function, the FA and IA have also been described to help build up rhythmic patterns and may also have pragmatic functions (e.g., IA potentially being used in an "emphatic" manner, see Di Cristo, 1999, 2000).

In English, however, an additional factor affecting metrical structure is lexical stress, which is an abstract representation and a distinguishing property of the word (e.g., the difference in meaning between the words *REcord*, the noun, and *reCORD*, the verb, with capital letters indicating stress). Consequently, lexically stressed syllables should be assigned to strong nodes. Metrical structure in English is also constrained by rhythmicality (the alternation of weak and strong prominences at each level in the hierarchy). Speakers may vary the metrical structure in order to avoid "stress clash" (two subsequently stressed syllables in a prosodic phrase, e.g., Shattuck-Hufnagel et al., 1994). Alternatively, speakers may insert a rhythmic stress on a normally unstressed syllable to avoid long stretches of speech without stress. Finally, English is said to generally prefer trochaic patterns within the level of the foot (where a foot generally consists of a strong syllable followed by a weak one), however at higher levels in the hierarchy it is said to be "right branching" in that stronger prominences tend to go to the right. (i.e.,

the Nuclear Stress Rule, Selkirk, 1984, Calhoun, 2010b for an overview). As a result, words may contain primary and secondary stresses which act as tonal targets for pitch accentuation in English. Thus, stress at both the lexical and phrasal level is cumulative, and pitch accents are considered to be *prominence-lending* at all levels of prosodic structure (i.e., word, phrase, etc.). This is in contrast with French, where lexical stress is not cumulative, causing pitch accentuation to function only on a phrasal level, and are *demarcative* in function (particularly FA), signaling the edges of the AP rather than lending the words or phrases as prominent.

To better illustrate the differences in prosodic structure between French and English, **Figure 1.5** shows the ToBI annotations for two utterances. The upper panel shows the prosodic structure of the French utterance "Moi, je décide d'aller au restaurant, elle préfère qu'on aille se reposer au cinéma", (*Me, I decide to go to the restaurant, she prefers that we relax at the cinema*) where the entire utterance occurs as one intonational phrase (bottom tier), divided into two intermediate phrases (third tier). The first ip is subsequently divided into three APs and the second ip contains five. The tonal targets shown on the second tier indicate at least one FA at the right edge of each AP, and IAs being produced on the left edge of the third, seventh, and eight APs, and coded as Hi. A high phrase accent is also indicated at the right edge of the first ip, and a high boundary tone at the right edge of the IP. The lower panel shows the prosodic structure of the English utterance "And there's lots of cases where we have more than one work mashed together," where the entire utterance occurs as one intonational phrase (bottom

tier), divided into three intermediate phrases (third tier). The tonal targets indicated on the second tier indicate at least one pitch accent indicating the prominent syllable within each ip, phrase accents at the right edge of each ip, and a boundary tone at the right edge of the IP. To sum up, French has three levels of phrasing, where the right and (optionally) left edge of the AP is marked with a pitch accent (FA or IA, respectively). This contrasts with English where there are only two levels of prosodic phrasing, and pitch accent gives prominence at all levels of prosodic structure.

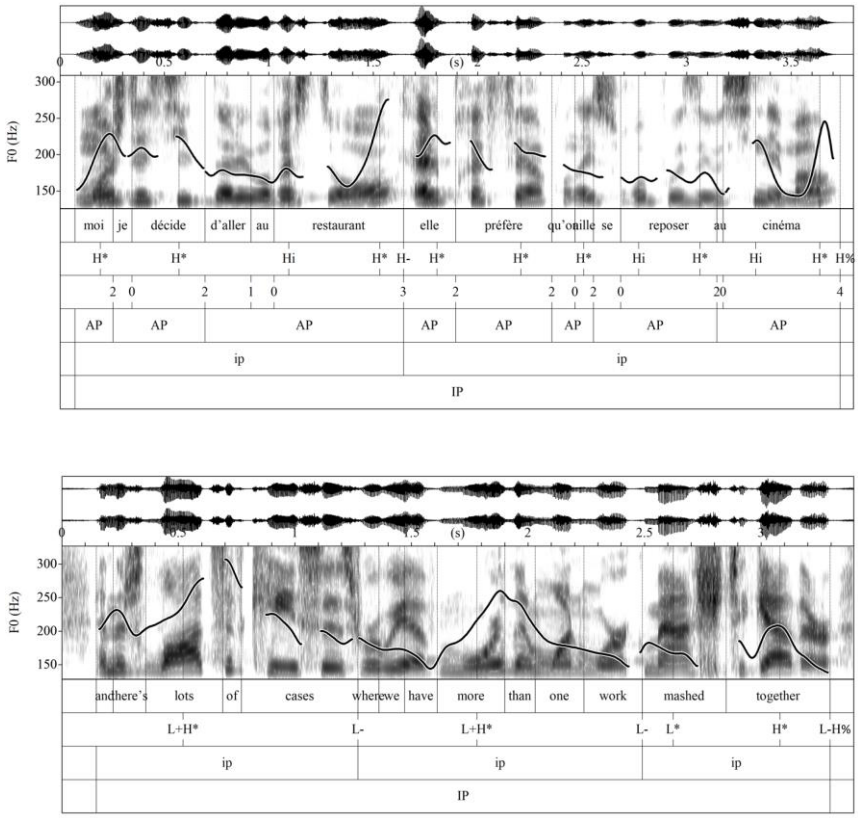

Figure 1.5 Differences in prosodic structure between French and English. **Upper panel:** The F_ToBI annotation of a French utterance taken from the French M3D-TED corpus by speaker JP (TEDx Talks, 2018) at 02:50. **Lower panel:** The MAE-ToBI annotation of an English utterance taken from the English M3D-TED corpus by speaker MS (Stewart, 2010) at 04:44.[1]

Within that AM approach, metrical structure continues to be relevant for higher levels of prosodic structure (e.g., the grouping of prosodic words in prosodic phrases), particularly in terms of pitch accent nuclearity. The concept of *nuclearity* refers to the position of a phenomenon within the prosodic structure. Specifically in the AM

[1] All Praat images were created using the Praat script developed by Elvira García (2017).

framework, *nuclear* pitch accents have been defined as the final (right-most) pitch accent within a prosodic phrase (e.g., Ladd, 2008). Following the right-branching bias laid out in metrical phonology, the nuclear pitch accent refers to the node in the metrical tree that is entirely dominated by strong nodes. Any strong syllables to the left of the nuclear pitch accent may receive a *prenuclear* pitch accent, and strong syllables to the right of the nuclear pitch accent may be stressed, but are generally much less acoustically prominent (see, e.g., Calhoun, 2010b; Ladd, 2008). The status of the nuclear pitch accent being on the node dominated by strong nodes makes it the most *structurally prominent* syllable. However, depending on the phonetic realization of pitch accents, prenuclear pitch accents may be perceived as being more *phonetically prominent* than nuclear pitch accents (e.g., see Ayers, 1996; Calhoun 2010b and references therein). Thus, following the aforementioned studies, the conceptualization of nuclear pitch accent adopted in this thesis is structural, considering the final pitch accent within the phrase as being nuclear, regardless of its relative phonetic prominence to prenuclear pitch accents. Finally, this concept of nuclearity can apply to pitch accents at both levels of prosodic phrasing (i.e., ip-nuclear pitch accent vs. IP-nuclear pitch accent, e.g., "more" vs. "together" in lower panel of **Figure 1.5**) and can even be applied to the relationship between the two phrasal levels (e.g., a nuclear ip, "mashed together" in lower panel of **Figure 1.5**).

While the AM model focuses largely on the realization of intonational contours which reflect prominence and phrasing, the

notion of speech rhythm is more related to metrical phonology. Defining and measuring speech rhythm is a challenging task, and a number of definitions have been proposed (see Turk & Shattuck-Hufnagel, 2013 for a review). At a basic level, we could consider rhythm as the temporal organization of prominence in speech (e.g., Astésano, 2001; Di Cristo, 2000). As such, rhythm is largely "situated at the interface between meter and surface constraints, be they prosodic (e.g., constituent size, rhythmic rules) or structural (syntactic, semantic, or informational)" (Astésano, 2017, p. 71; personal translation). Importantly for **Chapter 3** of this thesis, the domain of the AP (marked regularly by the FA) has been shown to be key in French rhythm, as it entails the (more or less) regular alternation between accented and unaccented syllables (e.g., Astésano, 2001; Delais-Roussarie, 1995; Pasdeloup, 1990).

The current subsection has offered a brief review of the prosodic structure differences in French and English. To summarize, speech in both languages is segmented into prosodic phrases at various levels, which each contain at least one phrasal prominence encoded in intonation as a pitch accent. In English, pitch accents generally associate with primary lexical stresses which are a property of the word and with a prominence-lending function, whereas in French, pitch accents associate in fixed (final and optionally initial) positions within the smallest prosodic phrase (the AP), having mostly a demarcative function. In the case of French, pitch accentuation is marking prosodic AP phrasing and thus the two prosodic units (pitch accentuation and phrasing) share a closer relationship than in English. Second, speech rhythm in French

functions on the domain of the AP, as it represents the regular alternance between prominent and less-prominent syllables. This is not the case for English, where pitch accents do not regularly mark prosodic edges. These two typological differences between French and English allow us to further assess the phonological synchrony rule with languages with different prosodic typologies, as most studies have focused on languages where pitch accentuation has a prominence-lending function, and fewer studies have assessed languages where pitch accentuation operates on a phrasal level. The following subsections will review the previous literature on the temporal association between gestures and speech prosody in terms of prosodic prominence, phrasal position, and rhythm.

1.3.2. Gesture and its temporal association with prosodic structure

1.3.2.1. Gesture and its temporal association with prosodic prominence

Since the description of the phonological synchrony rule (McNeill, 1992), a number of studies have empirically investigated the relationship between gesture and speech. While studies have used a variety of methods to investigate the issue, most have reported a close temporal association between gesture prominence (e.g., gesture strokes or apexes including various articulators such as the hands, head movements and eyebrow movements) and prosodic prominence (e.g., stressed syllables, pitch accentuation, etc.). The

current subsection will describe the findings of those studies while also commenting on the methodological differences between them.

One of the first studies to assess the phonological synchrony rule with empirical data was that by Nobe (1996). Using McNeill's own data, which was collected using a narrative retelling task in English, he chose a total of forty-eight representational (i.e., iconic or metaphoric) gestures to assess their relationship with prosodic prominence. Specifically, he used acoustic analyses to identify prominent syllables, which could contain the peak *f0* of the Intonational phrase, and/or the peak intensity of the intonational phrase - thus a single prominent syllable could contain both peak *f0* and peak intensity, or the two peaks may have occurred in two different prominent syllables. He found that in general, over 95% of the gestures coincided with a prominent syllable. Specifically, of the forty-eight gestures, thirty-six (75%) coincided with a syllable that contained both cues to prominence, while six gestures (12.5%) coincided with the syllable containing the *f0* peak, and three gestures (6.3%) with the syllable containing a peak in intensity. Based on these results, he proposed the rule of acoustic peak synchrony, suggesting that the prosodic cue encoding prominence may impact gesture production.

Another study that used the stroke of the gesture to assess its temporal overlap with speech prominence was by Karpiński et al. (2009). They studied task-oriented dialogues produced by Polish speakers. The authors used the Rhythm and Prominence (RaP) labeling system for prosodic annotation (Dilley & Brown, 2005),

where annotators perceptually identify weak and strong prosodically prominent syllables. In terms of gesture, they did not distinguish between gesture types. They found that 96% of the 223 gesture strokes overlapped with a prosodically prominent syllable, and that 75% of gesture strokes overlapped with a *strong* prosodic prominence. Another study by Shattuck-Hufnagel & Ren (2018) investigated the temporal overlap of referential and non-referential gesture strokes and ToBI-defined pitch accented syllables by an English speaker giving a 30-minute academic lecture. The authors found that 83.12% of gesture strokes overlapped with a pitch accented syllable. Specifically in terms of gesture referentiality, they found that 83.13% of non-referential strokes overlapped with pitch accented syllables, with similar rates for referential gestures (82.85%).

A number of studies have looked toward the apex (i.e., the "peak" of the stroke, see **subsection 1.2.3.**) to assess the temporal association between gesture and speech prosody. Loehr (2004; see also Loehr, 2012) analyzed four video clips extracted from a larger corpus of spontaneous dyadic conversation where speakers were merely asked to converse freely (the four video clips totaling 164 seconds for analysis). He annotated the gesture phases and apexes for manual gesture, as well as their type (following McNeill's classic gesture classification). Prosodic annotations followed the ToBI system. A qualitative analysis of his data showed that pitch accents and gesture apexes "co-occur repeatedly" (2004, p. 114). He then assessed the time distances between gesture apexes and their nearest accent, finding that their distribution is centered very close

to zero, and that the overall average is +17 ms (SD: ± 341 ms). Finally, using a time-window of 275 ms, he found that apexes and pitch accents tend to co-occur significantly more than other movement types or tone types. Furthermore, he found no differences between gesture types in terms of their relationship to speech prosody. Following a similar methodology, Jannedy & Mendoza-Denton (2005) found that in a spontaneous discourse by a single English speaker filmed at a town hall (totaling 130 seconds for analysis), 95.7% of the speaker's gesture apexes "co-occurred along with a pitch accent". Conversely, only 69.4% of the speaker's pitch accents were marked with a gesture apex. Importantly, the authors never specify the domain within which the two phenomena (i.e., apexes and pitch accents) co-occur, be they the bounds of the pitch accented syllables or an arbitrary time-frame following Loehr (2004).

Homologous to apexes, some studies have investigated what they term "hits." Gestural hits are defined as "an abrupt stop or pause in movement, which breaks the flow of the gesture during which it occurs. Hits appear as bouncing, jerky movements, changes in the direction of movement, or as complete stops in movement" (Yasinnik et al., 2004, p. 98). In a sample of 7.5 minutes of an academic lecture by an American English speaker, Yasinnik et al., (2004) investigated the co-occurrence of ToBI pitch accented syllables and gestural hits. They found a total of 130 polysyllabic words that co-occurred with a gestural hit. In 117 of these words (90%), the word also contained a pitch accent. This rate was much lower for monosyllabic words, where only 65% of the 116 hit-

aligned words were also pitch accented. Moreover, the authors mentioned that of the hit-aligned words that were not pitch accented, most were within 100 ms of a pitch accented word. Another study by Esposito et al. (2007) analyzed two monologues produced by two native Italian speakers instructed to speak freely about any topic (totaling over eight minutes) and found that gestural hits occurred within pitch accented syllables 78% and 84% of the time, respectively.

Supplementing research on natural speech, a number of experimental studies have also been conducted and have generally found a tight temporal association between apexes and pitch accentuation. For example, Leonard & Cummins (2010) asked one subject to read three fables two times each, where each fable contained three naturally stressed words where the reader was explicitly instructed to produce a beat gesture. They found that the closest speech landmark to the apex was the peak of the pitch accent in the stressed syllable. Esteve-Gibert & Prieto (2013) investigated deictic gestures produced by 15 native Catalan-speaking adults in a pointing-naming task. They found that the timing of the apex and the pitch peak was significantly correlated (for similar results incorporating oral articulatory gestures, e.g., articulatory movements of the lips, tongue, and throat for speech production, see Krivokapić et al., 2016; Rochet-Capellan et al., 2008; Roustan & Dohen, 2010). However, positive results have not always been found (e.g., see De Ruiter, 1998, Study 1, as cited by Esteve-Gibert & Prieto, 2013; Rusiewicz, 2010 for conflicting results). Recent lines of research have also begun measuring kinematic data through

motion tracking systems and have found close relationships between peak acceleration and deceleration and pitch peaks (e.g., Pouw & Dixon, 2019b)

Moreover, the study by Esteve-Gibert & Prieto (2013) found that not only prosodic (i.e., tonal) movements, but also gestural movements, are constrained by prosodic phrasing. Specifically, it is well-known that when a rising (L+H*) pitch accented syllable is in a phrase-final position before an upcoming low boundary tone, the *f0* peak is systematically shifted to the left. This allows for both the pitch-accent and the upcoming boundary tone to be produced within the tone-bearing unit. By manipulating the metrical structure of the target word that is produced along with a deictic gesture, they found that when the prominent syllable was phrase final, both the *f0* peak as well as the apex of the pointing gesture were shifted to earlier positions within the stressed syllable. The results thus suggest that gesture apexes behave much like intonation peaks in that they are both constrained (either directly or indirectly) by prosodic phrasing. In addition to being bound by prosodic structure, another study by Krivokapić et al. (2017) showed how prosodic structure affects gesture production. The authors asked two native English-speaking participants to perform a pointing-naming task while recording their movements with a motion capture system. They found that deictic gestures showed lengthening (specifically in the return movement of pointing) under prominence as well as at ip phrase-final positions. Taken together, the results of these latter two studies suggest that not only prominence, but phrase-level prosodic structure indeed impacts the realization of co-speech gesture.

The aforementioned studies have focused on manual co-speech gestures, but similar findings have been found when looking at non-manual gestures (e.g., head and eyebrow movements). For example, Alexanderson et al. (2013) found that in Swedish spontaneous speech, head nods are closely associated with stressed syllables (particularly those containing a focal accent), and that the apex of the nod was on average aligned with the nucleus of the stressed syllable. Esteve-Gibert et al. (2017), investigated head gestures in semi-spontaneous speech by 24 native Catalan speakers and found that the apexes of head gestures were synchronized with pitch accented syllables. Moreover, they found that head nods were also bound by prosodic structure in a similar fashion to deictic gestures (Esteve-Gibert et al., 2013).

The relationship with eyebrow movements is less clear. Some studies have found eyebrow movements to associate with *f0* movements and accented syllables (Cavé et al., 1996; Guaïtella et al., 2009; Flecha-García, 2010). However, the relationship between eyebrow movements and *f0* movement is not is not as straightforward, for example Swerts & Krahmer (2010) found that in a corpus of newsreaders (4 speakers, with a total of 60 sentences full sentences selected for analysis) strong accents are indeed accompanied by eyebrow movement (more so than "no accent" or "weak accents"), but conversely found that many eyebrow movements are produced without an accompanying strong accent (see also Berger & Zellers, 2022). However, when studying the combination of (focal) pitch accentuation, head movements, and eyebrow movements, Ambrazaitis & House (2017) found that in a

corpus of Swedish news readings (31 brief news readings by four speakers, totaling over six minutes of speech), focal pitch accents tended to associate with a head movement alone, or both a head and eyebrow movement together. Least present in their data was the coupling of pitch accent and eyebrow movement alone. The authors suggest that eyebrow movements may have less autonomy than head movements and do not have a prominence-lending function on their own, but rather couple with pitch accents and head movements together to intensify the prominence-lending effect of head nods. Evidence from more recent studies assessing the size of the eyebrow movement show that larger *f0* excursions correlate with larger eyebrow movements (e.g., Berger & Zellers, 2022) and that *f0* rises become increasingly larger when produced with additional gestures (i.e., a head movement and a head + eyebrow movement, Ambrazaitis & House, 2022) suggesting a cumulative-cue hypothesis, where "the acoustic realization of pitch accents … co-varies with the number of accompanying gestures … and that this cumulative relation might be to some degree sensitive for lexical prosody." (Ambrazaitis & House, 2020, p. 26). In other words, prominence in speech and prominence in gesture seem to go hand in hand (see also Krahmer & Swerts, 2007 for the effects of manual gesture production on the acoustic realization and perception of speech prominence).

1.3.2.2. Gesture and its temporal association with prosodic phrasing

As mentioned before, while the majority of studies have assessed the temporal alignment between gestures and prosodic prominence,

a smaller number of studies have also investigated alignment of gestures with prosodic phrases. A few studies have specifically focused on the production of G-Phrases (see **subsection 1.2.3.**) and their relationship with the intermediate phrase. For example, Loehr (2004, 2012) found that the onset of G-Phrases generally coincide with the onset of ips over two thirds of the time, and that the general tendency is for the G-Phrase to slightly precede its corresponding ip (average lead time of three frames, or 100 ms). The author also describes how often multiple G-Phrases were found to occur within a single ip. In such cases, the number of G-Phrases never exceeded 3, and they were often aligned with a syntactic constituent, or a slight pause that was not worthy of a phrasal boundary. However, studies in other languages such as French (Ferré, 2010), Italian (Cantalini & Moneglia., 2020), Turkish (Turk, 2020), and Brazilian Portuguese (Barros, 2021) have found that generally the duration of the G-Phrase spans the entire ip. In other words, the onset of the G-Phrase precedes the onset of the ip, and the offset of the G-Phrase occurs after the offset of the ip. Furthermore, Karpiński et al. (2009) found no consistent results in terms of alignment with G-Phrases and ips in Polish.

Investigating higher levels of prosodic phrasing, Yasinnik et al. (2004) assessed the temporal association between groups of IPs and G-Units, namely by assessing pause durations between IPs. Their hypothesis was that multiple IPs could group together to form higher-level prosodic constituents, where the pause at the end of the constituent would be larger than any within-constituent pause, which may correspond to the utterance level as proposed by Selkirk

(1978) among others (see Shattuck-Hufnagel & Turk, 1996 for a review), and that these would correspond to groupings of gesture phrases (similar to G-Units as defined in **subsection 1.2.3.**). Indeed, they found that 12 of the 14 cross-IP gesture groupings fell within one of these higher-level prosodic constituents. Following up on these initial results, Shattuck-Hufnagel & Ren (2018) assessed the association between perceived gestural groupings (PGGs) and higher-level prosodic constituents. The authors identified PGGs by assessing form, so that gestures that were perceived as having similar kinematic characteristics were grouped together to form a PGG. Regarding the identification of the higher-level prosodic constituents, the authors carried out an extended version of Rapid Prosodic Transcription (e-RPT, see Cole & Shattuck-Hufnagel, 2016) where eight participants with no training in prosodic annotation listened to speech and assessed three levels of perceived boundary strength by putting between one and three slashes. Participant markers of boundary strength are then summed across listeners to provide an estimate of higher-level prosodic groupings. The authors used an arbitrary cumulative number of 15 annotation marks as a threshold to identify these higher-level prosodic constituents. As a result, the e-RPT annotations returned a total of eight higher-level constituents, made up of 66 IPs and 100 ips. Their observations suggest that the PGGs seem to roughly align with these higher-level constituents in six out of the eight cases. However, the authors concede that such a methodology is rather preliminary and the results should be taken as rather suggestive and aim to open future lines of study.

1.3.2.3. Methodological differences between studies

It is important to highlight four major methodological differences between the aforementioned studies. First, the choice of gestural landmark varies across studies. The second is that studies have varied greatly in the type of gesture under investigation, with few studies actually comparing the alignment of different gesture types. A third issue is the type of speech that is analyzed, with studies largely investigating either naturally produced speech or controlled speech where participants are explicitly instructed to gesture on particular words. Finally, the alignment criteria is an important methodological choice for the interpretation of the results. Taken together, it can be noted that experimental studies with controlled speech (i.e., speakers are explicitly instructed to produce gestures on words that are specifically elicited to also contain a pitch accent) have used more continuous variables (e.g., time distance from the gestural and prosodic landmarks) to assess synchrony, which is then qualitatively interpreted to show co-occurrence. Studies on naturally produced speech, however, have often used more categorical assessments (occurrence or non-occurrence within a set time frame). Importantly, studies that have assessed the apex as a gestural landmark have generally used relatively arbitrary time frames to interpret co-occurrence (e.g., average syllable duration, as in Turk, 2020; 275 ms in Loehr, 2004, 2012), which does not precisely assess whether apexes are actually occurring within the bounds of a specific prosodic unit (e.g., the pitch accented syllable). In order to be able to compare our results with the preceding literature, a comprehensive analysis will be carried out in **Chapters**

3 and 4 of the current thesis, investigating the alignment patterns of both strokes and apexes of both referential and non-referential gestures using more explicit alignment criteria (namely, stroke overlap with a pitch accented syllable, and the occurrence of the apex within the bounds of a pitch accented syllable) to more precisely assess the integration of gesture production with speech prosody in terms of pitch accentuation.

1.3.2.4. Refining the phonological synchrony rule: Typological differences and nuclearity

Methodological differences aside, the aforementioned paragraphs have given a general overview of the studies that show how speech prosody and gesture tend to go hand-in-hand, generally finding that gesture is attracted to prominence in speech and is also affected by aspects of higher-level prosodic structure. However, most of the studies have focused on languages where pitch accentuation is prominence-lending. Fewer studies have investigated the temporal association between gesture and speech where intonation is not considered prominence-lending. For example, Fung & Mok (2018) investigated gesture-speech synchrony in Hong Kong Cantonese (a tonal language) using a pointing-naming task that induced contrastive focus. The authors found that *f0* did not encode prosodic prominence in such conditions, but rather prominence was encoded through durational lengthening on focused syllables. Moreover, they found that *f0* did not play a role in gesture-speech synchronization, but rather the word carrying prosodic stress acted as the anchor point, with apexes regularly occurring on the first

syllable regardless of whether it carried prosodic stress or not. Interestingly, some recent studies have begun assessing gesture-speech synchrony in order to better understand prosodic structure in languages where prosodic models are under debate. For example, Turk (2020) investigated the association between tonal movements and gesture in a corpus of narrative retellings by four native Turkish speakers to assess the status of a phrase-final pitch rise as either a pitch accent or a boundary tone. The author found that gestures generally did not associate with the final rise, but rather with a L tone at the left edge of the prosodic word. The authors thus interpreted this as suggesting that the phrase final rise was indeed a boundary tone (see also Kaufman & Farinella, 2022 for Malay). Thus **Chapter 3** of this thesis will assess gesture-speech synchrony in French, a language where pitch accentuation has been shown to have a demarcative rather than prominence-lending function (see **subsection 1.3.1.**). The results of which will clarify not only how gesture and speech are associated in French, but may also help disambiguate the status of the IA and clarify the role of prosodic edges in the attraction of gesture.

In addition to the need to study languages in which pitch accentuation is not prominence-lending, much less is known about the role of phrasal position (i.e., pitch accent nuclearity) in gesture-speech synchrony. Only one study to our knowledge has specifically investigated this issue. Using data from two dyadic spontaneous conversations, McClave (1998) investigated whether referential gesture strokes occurred with stressed syllables, and particularly whether they tended to be the nucleus of the Tone Unit

(as defined by Cruttenden, 1986, cited in McClave, 1998). She defines the nucleus as "the last stressed syllable with a significant change in pitch" (p. 84). She found that in 53% of the Tone Units which contained referential gestures, the stroke co-occurred with the nucleus. An additional 25% of tone units contained strokes which co-occurred with stressed syllables which were not acting as the nucleus. Importantly, the author did not clarify whether the syllables assessed were pitch accented or not (merely describing them as stressed or acting as the nuclear stress). Furthermore, she did not control for cases in which there was only one prominence in the Tone Unit, thus the results that gestures are associating more with nuclear prominence could be merely a byproduct of the relative frequency of nuclear prominences compared to prenuclear ones. In other words, in cases in which tone units contain only one (nuclear) pitch accent, the gesture would naturally associate with the only prominent position in the phrase, biasing the general results. Moreover, even though the author describes how the nuclear stress is not always the most prominent in the Tone Unit, she limits her analysis to the nuclear/prenuclear distinction.

1.3.2.5. Gesture and its temporal association with rhythm

Additionally, only a handful of studies have investigated the rhythmic production of subsequent gestures (that is, groups of subsequent gestures that are perceived to be "beating musical time"). For example, McNeill (1992, p. 244) found that gesture strokes tend to occur at more-or-less equal intervals of one or two seconds, depending on the speaker. A study by McClave (1994)

found that subsequent beat gestures tended to be more isochronically produced than subsequent referential ones. Moreover, she found that groups of subsequent beat gestures were produced such that one gesture within the group associates with the nuclear pitch accents, and the rest of the gestures span out from this anchoring point following their own tempo independent from that of pitch accentuation, causing some gestures to associate with prominence in speech, while others do not. Similarly, Loehr (2007) found similar findings for manual gestures, and also included other articulators such as head movements and blinks. He found that each articulator follows its own rhythm independent of speech rhythm. In French, however, pitch accentuation and phrasing share a closer relationship, and the bounds of the AP have been shown to be key in speech rhythm. Thus, the second aim of **Chapter 3** of this thesis will be to assess whether this key rhythmic role of the AP also translates to the visual domain in French, and to compare potential differences by gesture type.

To sum up, there is a current need to refine McNeill's phonological synchrony rule and empirically assess how gesture and speech are synchronized on a phonological level not only considering prosodic prominence but also complementary prosodic features like prosodic phrasing and phrasal position/nuclearity, both key aspects of complex prosodic structure. Such an approach is likely to be key to be able to disentangle the effects of relative prominence and pitch accent nuclearity in attracting gesture production. For these reasons, **Chapter 4** of this thesis will aim to better understand the effects of pitch accent nuclearity in English TED Talks by specifically

controlling for the number of potential prosodic anchoring points available, as well as including an analysis of relative prominence to assess whether phrasal position has an effect on gesture-speech production.

Methodologically, the two empirical studies will systematically assess the phonological behavior of different manual gesture types (i.e., referential and non-referential gestures) and more specifically, the temporal landmarks investigated will include both the gesture stroke and the apex of the corresponding manual gestures. As we have reviewed in **subsections 1.3.2.1.** and **1.3.2.3.**, previous studies on the temporal association between gesture and pitch accentuation have assessed one of the two gestural landmarks, but very few have assessed the two by accounting for gesture type at the same time. Even though studies regarding the apex have mostly found tight association in laboratory studies between gesture apexes and pitch peak, many studies using natural speech have used rather wide margins to consider their temporal co-occurrence. This is why it is important to assess the association patterns of these two gestural landmarks, while also accounting for gesture type.

All in all, the current subsection has reviewed the previous literature on how gesture and prosodic structure are temporally associated. **Chapters 3 and 4** of this thesis will help refine this relationship by systematically taking phrasal prosodic structure into account. Furthermore, it has already been claimed how gesture and speech convey the same pragmatic meaning (e.g., McNeill's 1992 pragmatic synchrony rule). This has even been shown to be the case

when specifically looking at the pragmatic relationship between gesture and speech prosody (namely, signaling completeness, emphasis, or aspects of information structure, see Loehr, 2012). Even though a review and tentative proposal of the different pragmatic functions of gesture will be presented in the context of M3D (see **Chapter 2** for a short description), more empirical studies are needed to assess the various pragmatic functions of gestures while adopting a more systematic and standard approach. **Chapter 5** of this thesis will specifically investigate one such pragmatic function: the marking of information structure. The upcoming **subsection 1.4.** will thus offer an overview of what information structure is, and will review the different prosodic and gestural cues that speakers use to mark information structure in speech.

1.4. Multimodal cues to information structure

Gesture and speech have been claimed to be pragmatically coherent, as per the pragmatic synchrony rule (McNeill, 1992). While numerous studies have brought to light a number of pragmatic functions of gestures, only a handful have assessed the pragmatic role of speech and gesture in a more systematic manner. In the context of the present thesis, two main objectives will deal with the pragmatic functions of gesture. First, one of the goals of the new M3D proposal in **Chapter 2** will be to review previous work on the pragmatic functions of gestures across a number of subfields within the field of pragmatics, aiming at adopting standard practices in the field for assessing pragmatics. Second, we will systematically

assess the role of prosodic and gestural cues in the marking of one particular aspect of pragmatics: Information structure. The following subsections will describe the theoretical background and the multimodal cues to information structure.

1.4.1. General overview of information structure

Information structure (henceforth, IS) can be generally described as how speakers "package information" in speech for their interlocutors in order to update the shared common knowledge between them (i.e., their common ground), ultimately moving communication forward (Chafe, 1976, as cited in Krifka, 2008). IS can be seen as a part of discourse structure, and studies generally distinguish between information which is *old* and information which is *new*. While previous literature has used various terms that often overlap in meaning (see, e.g., Krifka, 2008; Skopeteas et al., 2006 for a discussion), IS can generally be interpreted through three independent, non-mutually exclusive dimensions: focus (as opposed to background), topic (as opposed to comment), and the information status of referents (Ritz et al., 2008).

Focus can be defined as "the presence of alternatives that are relevant for the interpretation of linguistic expressions" (Krifka, 2008, p. 247), and is often seen as the response to an (implicit or explicit) wh- question. Focus can also be considered as new information that helps move discourse forward (e.g., Götze et al, 2007). As such, old information would correspond to information that is already in the common ground of the speakers and can be

called *background*. **Example 1.1** below shows a question-answer pair where the focused element clearly indicates the potential for alternatives, as someone else (e.g., John, Mary) may have stolen the cookies. In addition to single words acting as focus in response to a question (i.e., *narrow* focus as per Ladd, 1980), entire clauses may be considered focused (also known as *broad* focus, see **Example 1.2**). Importantly, all utterances generally contain at least one focused element, while background information is optional.

[1.1] Q: *Who stole the cookies?*
A: [*Peter*]$_{\text{Focus}}$ [*stole the cookies.*]$_{\text{Background}}$

Taken from Krifka (2008, p. 250)

[1.2] Q: *What's that noise?*
A: [*Our neighbors are renovating.*]$_{\text{Focus}}$

Taken from Arnhold et al. (2016, p. 2)

Another set of terms that are used in a similar manner to focus and background are the terms *topic* and *comment*. We use the term *aboutness topic* as described in Féry (2017) to indicate referents about which the remainder of the sentence is predicated, or "commented on", where the predication (or *comment*) typically contains a focused constituent. **Example 1.3** shows how by changing the question that was asked in Example 1, we can interpret the phrase as *Peter* being the (aboutness) topic of the sentence and the remainder the comment regarding what Peter did. Topics can be distinguished as *frame-setting* topics, where the topic is constituted

of adverbial expressions which "set the frame in which the following expression should be interpreted" (Krifka, 2008, p. 278, **Example 1.4**), limiting the interpretation to a specific domain. As Krifka (2008) mentions, it is very easy to confuse focus/background with topic/comment, as topics tend to deal with old information and comments with new information (as in **Example 1.3**, where comment and focus constituents coincide). However, topics can also introduce discourse-new entities (**Example 1.5**), and comments do not align perfectly with focused constituents (**Example 1.6**).

[1.3] Q: *What did Peter do?*
A: [*Peter*]$_{\text{Aboutness Topic}}$ [*stole the cookies.*]$_{\text{Comment}}$

Adapted from Krifka (2008, p. 253)

[1.4] Q: *How is John?*
A: [*Healthwise*]$_{\text{Frame-setting Topic}}$ *he is fine.*

Taken from Krifka (2008, p. 278)

[1.5] [*A good friend of mine*]$_{\text{Topic/New}}$ [*married Britney Spears last year*]$_{\text{Comment}}$

Taken from Krifka, 2008 (p. 273)

[1.6] Q: *When did* [*Aristotle Onassis*]$_{\text{Topic}}$ *marry Jacqueline Kennedy?*

A: [*He*]$_{\text{Topic}}$ [*married her* [*in 1968*]$_{\text{Focus}}$]]$_{\text{Comment}}$

Taken from Krifka (2008, p. 273)

Finally, individual discourse referents can be identified as new, accessible, or given (e.g., Götze et al., 2007, among others). This has been referred to in the literature as the *Information Status of Referents* (henceforth, ISR), and relates to the degree to which discourse referents (i.e., noun phrases or prepositional phrases that specifically refer to entities in discourse) are cognitively active for the addressee (Chafe, 1974, as cited in Krifka, 2008). *Given* referents have been explicitly mentioned in previous discourse and are thus cognitively active. Less cognitively active referents are said to be *accessible*, in that they can be situationally or contextually inferred, or are assumed to be familiar to the addressee through general cultural or world knowledge. These may also include unique referents, such as *the Sun* or *Barcelona*. Accessible referents can be further classified by their mode of accessibility (i.e., whether they can be textually accessible, situationally accessible, or inferentially accessible), and particularly for the latter, by the relationship they have with antecedents, for example, part-whole (hand - finger), set relationships (subset/superset/same set, e.g. flower - lily), or entity attribute (flower - their scent), see Baumann & Grice (2006) and Götze et al. (2007). Finally, *new* referents are those which have not been used in context and are thus cognitively inactive for listeners (see **Example 1.7** for each type, adapted from

Götze et al., 2007). Much like the other two dimensions, and as briefly mentioned before, these are not mutually-exclusive: new referents may be found in the topic and given referents may be marked with focus (see, e.g., Ambrazaitis, 2009, p. 21 and references therein).

[1.7] [*Peter*]$_{\text{New}}$ *went to* [*the garden*]$_{\text{New}}$. [*The flowers*]$_{\text{Accessible}}$ *were blooming and* [*he*]$_{\text{Given}}$ *was happy.*

Adapted from Götze et al. (2007)

A final aspect of IS is *contrast*. Contrast is usually described in terms of focus insomuch as focus refers to "the presence of alternatives" (as per Krifka, 2008), and contrast refers to *explicit* alternatives from a limited set already in the discourse (e.g., Repp, 2010; see also Calhoun, 2009). Contrast can come in a variety of flavors (e.g., contrastive focus, corrective focus, **Examples 1.8** and **1.9** respectively).

[1.8] Q: *Did he move to the red house, or the blue house?*
A: *He moved into the* [*red*]$_{\text{Contrastive focus}}$ *house.*

[1.9] A1: *Peter stole the cookies.*
A2: *No,* [*Anne*]$_{\text{Corrective focus}}$ *stole the cookies.*

The status of contrast within notions of IS are debated, where some researchers maintain that contrast is a sub-type of focus (e.g., Büring, 1997; Krifka, 1998, 2008), which can then combine with topics to form "contrastive topics" (see **Example 1.10**), whereas

other researchers suggest that it may be an IS construct in its own right (e.g., Molnár, 2002; Repp, 2010), interpreting contrastive topics as "Topic + Contrast", a construct separate from focus.

[1.10] Q: *What do your siblings do?*

A: [*My* [*sister*]$_{\text{Focus}}$] $_{\text{Topic}}$ [*studies medicine*]$_{\textit{Focus}}$ *and*
[*my* [*brother*]$_{\textit{focus}}$]$_{\textit{Topic}}$ *is*
[*working on a freight ship.*]$_{\text{Focus}}$

Taken from Krifka (2008, p. 276)

While IS can be made apparent through syntactic, morphological, and lexical means, the most relevant markers of IS for this thesis are prosody and co-speech gesture. Interestingly, studies on the prosodic marking of IS have shown that regardless of prosodic typological differences, this prosodic variation can be subsumed by underlying principles, suggesting some common phonological structures (e.g., reduction of given information; for a review, see Kügler & Calhoun, 2020, and references therein). In terms of the gestural marking of IS, most studies have centered on the gestural marking of ISR, yet have found fairly consistent results in that gestural production is sensitive to IS (see Debreslioksa & Gullberg, 2022 for a review). However, only one study to our knowledge has assessed the use of gesture and prosodic prominence together as markers of IS, and none have explored the complex interaction that may arise between the two. **Subsection 1.4.2.** will briefly review the literature on the prosodic cues to IS, while **subsection 1.4.3.** will briefly review the literature on the gestural cues to IS.

1.4.2 Prosodic marking of information structure

Prosody has been shown to be a principal marker of IS in a variety of languages (see, Baumann, 2006; Kügler & Calhoun, 2020 for reviews). In languages where pitch accents tend to have a prominence lending function (e.g., English, German, Dutch), it is generally held that newer information in speech receives greater prosodic prominence, while given information receives less prominence. Studies on the prosodic marking of IS have largely centered on the marking of focus, finding that the focused word is the most prosodically prominent in an utterance.[2] However, two views regarding the relationship between prominence and focus have emerged. The *direct-relationship* view largely holds that the phonetic and phonological cues to prominence map directly onto associated meanings (that is, focus, e.g., Xu & Xu, 2005; Breen et al., 2010). However, the view that is most widely accepted within the AM model is that prominence *indirectly* marks focus through nuclear pitch accentuation. That is, phonological and phonetic cues to prominence *generally* map onto nuclear pitch accentuation, which is then the structure that is used to mark focus. In other words, the relationship between acoustic prominence and focus is mediated by phonological categories. Thus, according to this view, nuclear pitch accentuation is key in marking focus (e.g., Calhoun, 2010b; Ladd, 2008; Selkirk, 1995).

[2] This relationship is rather straightforward when the focus constituent is a single word (e.g., narrow or contrastive focus), but the relationship is less clear in broad focus conditions. This goes beyond the scope of the current thesis, but see Calhoun, 2010b for a discussion on focus projection.

Within the two views, the role of prenuclear accents and post-nuclear prominences remains a matter of debate in Germanic languages. While some researchers hold that prenuclear pitch accents are merely "ornamental" (Büring, 2007) or not reliable for focus marking, but rather markers of metrical stress which have important rhythmic functions (Calhoun, 2010a), some researchers have found evidence that they may indeed play a role in focus marking in some contexts. For example, prenuclear accents may be used to mark focus when sentences contain multiple foci (e.g., contrastive topics) (Calhoun, 2010b; Féry & Samek-Lodovici, 2006). While post-nuclear prominences are generally either deaccented (do not carry a pitch accent) or occur in a narrow pitch range (Kügler & Féry, 2017), full post-nuclear pitch accents may mark focus in particular discourse contexts, such as second occurrence focus (see Beaver et al., 2007; Baumann, 2016).

Fewer research has looked at the prosodic marking of topic/comment. Given the interactions between syntax and metrical/prosodic structure, it is generally held that topics tend to contain given information and are unaccented, while comments convey new (and prosodically marked) information. However, topics may indeed receive pitch accentuation (such as in cases of contrastive topics). Researchers have suggested that topics are then prosodically realized differently from focus. For example, Calhoun (2010b; 2012) holds that topics tend to be less prosodically prominent. Other studies have suggested that topics are generally produced with rising pitch accents (i.e., L+H*, L*+H), while foci

are produced with falling accents (H*+L) (Büring, 2003; Steedman, 2014 for English).

Finally, the ISR may also affect the prosodic realization of the referential expression. Given referents are generally associated with lower stress-based prominence and/or deaccentuation, while new referents are associated with perceptually stronger prominence and triggers accentuation (e.g., Baumann & Grice, 2006; Cruttenden, 2006; Ladd, 2008; see Kügler & Calhoun, 2020 for a review). However, this tendency may be affected by other dimensions of IS. For example, given referents that fall in the background of an utterance are more likely to be unaccented than new referents (Féry & Kügler, 2008; Gussenhoven, 1983; Selkirk, 1995) but may receive (relatively less prominent) pitch accents to comply with rhythmic constraints (e.g., Baumann et al., 2007; Calhoun, 2010b; Féry & Kügler, 2008). Additionally, focus marking may override referential givenness, so that when the given referent is in focus, it will receive a pitch accent (e.g., Baumann & Riester, 2012). A study by Braun (2006) showed that given referents acting as aboutness topics regularly receive pitch accentuation in German, but the phonetic realization of those accents differs when those elements are contrastive, with contrasted referents being produced with higher and later pitch peaks. While much of the literature has centered on the given/new distinction, some studies have shown that accessible referents are more variable in their association with prosodic cues, and may be highly dependent on the relationship with its antecedent in speech (Baumann & Grice, 2006).

The relationship between newness and prominence has been also described in terms of tone type. Some have suggested a near one-to-one (i.e., categorical) mapping between tone type and newness, where (in English) L+H* associates with new and/or contrastive elements, H* and !H* with non-contrastive new or accessible referents, and L* or no accent with given information (see e.g., Pierrehumbert & Hirschberg, 1990). Others have considered more gradient scales (e.g., Gussenhoven's "Effort Code," 2002) or a combination of categorical and gradient scales (e.g., Baumann et al., 2006; Calhoun, 2009). More recent views propose a probabilistic or distributional relationship between tonal categories and newness. Regarding focus types, Mücke & Grice (2014) found that while focus tended to be marked by the presence of pitch accents, there was not a categorical difference but rather a distributional difference in tone type, with L+H* being used increasingly from broad, to narrow, and contrastive focus. Similarly, Im et al. (2018) found that all tone types were found across ISR categories, with a tendency for given referents to be unaccented (for similar results, see Baumann & Riester, 2013).

All in all, the literature presented in this subsection suggests that, particularly for Germanic languages, newer information typically receives greater prominence, which may or may not map to phonological categories in terms of pitch accent type or nuclear status. The following subsection will describe how gesture associates with IS.

1.4.3. Gestural marking of information structure

A handful of studies have investigated how gesture associates with IS through the lens of the Communicative Dynamism approach (CD), which refers to the degree to which an utterance moves the discourse forward (Firbas, 1971). In line with Givón's (1983) Principle of Quantity (see also the theory of Effort Code by Gussenhoven, 2002), McNeill (1992, 2005) suggested that gesture production occurs as a function of CD. According to McNeill, a narrator's gesture reflects which elements of a story are the most crucial for advancing the story. Specifically, he predicts that speech with low CD (i.e., background, topic, and given referents) has less likelihood to co-occur with gesture, while speech with high CD (i.e., focus, comment, and new referents) are more likely to be co-produced with gesture. Furthermore, he proposed to match different gesture types to different levels of CD. Namely, that non-referential and pointing gestures will accompany speech with lower CD, and iconic and metaphoric gestures will accompany speech with higher CD, as according to Givón (1985), "the less predictable/accessible/continuous a topic is, the more coding material is used to represent it in the language" (p. 197, as cited in McNeill, 1992). Much research since has offered supplemental evidence to these claims.

At the level of focus (elements with high CD), a number of studies to our knowledge have empirically investigated the relationship between gestural and prosodic marking of focus and/or contrastive focus. A handful of studies have focused on non-manual gestural

cues to focus. For example, Ambrazaitis & House (2016, 2017) found that in a corpus of Swedish news reporting, head nods and eyebrow movements occurred most often with focal accents, which are phonetically realized with an additional pitch rise (also termed "big accent") and are said to mark sentence-level prominence and generally (though not always) correspond to focus. Similarly, head nods and eyebrow movements have been found to co-occur with contrastive focus in French (e.g., Dohen et al., 2006; Roustan & Dohen, 2010). Importantly, such movements aid in the perception of contrastive focus (e.g., Prieto et al., 2015). In terms of manual co-speech gesture, one study by Ebert et al. (2011) investigated how strokes align with nuclear pitch accentuation (which have been described to stably mark focus), and G-Phrases align with focus constituents in the Bielefeld SAGA corpus (Lücking, et al., 2013). For every G-Phrase identified in a 20-minute extract of the corpus, the authors added the corresponding information-structural annotations, namely by identifying nuclear pitch accents, focus constituents, separating new-information foci (i.e., broad and narrow focus) from contrastive foci. The authors found that of the 275 G-Phrases annotated, only 10 did not associate with a focus constituent. Looking at temporal relationships between G-Phrases and focus constituents, the authors found that of the 260 G-Phrases that associated with new-information foci, G-Phrase onsets coordinated closely with focus constituent onsets, starting an average of 310 ms (SD: 410 ms) before the onset of the focus constituent. Few cases were observed where focus begins before G-Phrase onset. The relationship between the end of the focus

constituent and gesture offset (i.e., the end of the stroke) were much more variable. The 56 G-Phrases that associated with contrastive foci tended to start earlier than with new-information focus (770 ms earlier on average), however there was also a much higher variability (SD: 700 ms). The authors suggest that gestural association with contrastive foci shows a more loose temporal relationship. Regardless of the precise temporal relationship, these results suggest that gestures tend to coincide with focus constituents.

Ferré (2014) aimed to understand the interaction between prosody, gesture, and syntactic marking of focus via marked structures (i.e., different forms of syntactic fronting). Indeed, in French, syntactic fronting is a common strategy to mark focus (e.g., "*Y avait ma soeur et des amis qui étaient venus me rejoindre*" (literally translated as "There were my sister and some friends who came to visit me;" example taken from Ferré, 2014, p. 270). Her study involved the analysis of a corpus of spontaneous French conversation by three pairs of speakers (total duration: 1h30). Different types of syntactic fronting, prosodic emphasis, and gestures were annotated. The results showed that while all three strategies (i.e., fronting, prosody, and their association with gesture) were used for marking focus, they were used in a complementary fashion (in so much that speakers generally do not mark focus in all three modes simultaneously). When focusing on manual gesture type, the author found that beat gestures co-occurred with prosodic focus marking much more than other gesture types,

and metaphoric gestures co-occurred with syntactic fronting more than other gesture types.

A handful studies have focused on the gestural marking of ISR, with a number of studies having found that gestures tend to mark the introduction of new entities in discourse (or the reintroduction of a given referent with a full noun phrase; see Debreslioksa et al., 2013; Debreslioska & Gullberg, 2019; Gullberg, 2003, 2006; Levy & Fowler, 2000; Marslen-Wilson et al., 1982; Yoshioka, 2008). However, given referents which are maintained from one clause to the next are generally produced with lexically reduced forms such as pronouns or zero anaphora and are oftentimes produced without gesture (e.g., Debreslioska et al., 2013). However, many of the abovementioned studies have only distinguished new referents from (maintained or reintroduced) given ones.

A recent study by Debreslioska & Gullberg (2020b) included accessible referents. The researchers found that in a narrative retelling task, accessible referents were generally encoded syntactically with definite noun phrases, while brand-new referents were encoded syntactically with indefinite nominal expressions/noun phrases (see also, Clark, 1975, 1977; Gundel, 1996; Prince, 1992). Specifically, accessible referents were significantly more likely to be marked by a gesture than new referents. While these findings seem to contradict the previous studies and McNeill's own suggestion about newer information being marked gesturally, the authors suggest that, in fact, speakers mark accessible referent because indeed they are linguistically

coded as given referents, but they should be considered new for the listener. To further assess whether it is indeed the "richness" of the referential expression's morphosyntactic form or the ISR, Debreslioska & Gullberg (in press) conducted an experimental study focusing on given referents, where participants had perform a narrative retelling task in three conditions: a control condition (no instructions), a noun condition (participants were instructed to rename every discourse referent with a full noun form) and a pronoun condition (participants were instructed to use only pronouns). Given referents were then categorized into three types. Reintroduced referential expressions were referents that were mentioned after a gap of one or more clauses and instantiated as the grammatical subject. Maintained SS referential expressions were those that were mentioned in the immediately preceding clause and were instantiated as grammatical subject, while maintained OS were those instantiated as the grammatical object. They found that gestures were more likely to be produced with referential expressions in the noun condition than the pronoun condition. Comparing within the control condition, they found that referential form (noun vs. pronoun) was a greater predictor of gesture production, whereas in the noun and pronoun conditions, ISR was a better predictor, occurring significantly more with reintroduced referents than maintained ones (with referents instantiated as grammatical objects receiving significantly more gestures than those instantiated as grammatical subjects). All in all, these studies show a tight relationship between ISR, morphosyntactic form, and gesture production.

Further studies have found that gesture form is modulated by ISR, in that gestures which co-occur with new referents tend to be more complex character-viewpoint gestures that encode entity information such as size or shape, while gestures which co-occur with given referents tend to be less complex observer-viewpoint gestures that encode action information (Debreslioska & Gullberg, 2019, 2020a; Foraker, 2011, as cited in Debreslioska & Gullberg, 2020a). Moreover, when a gesture denoting a referent is repeated (i.e., both the referent and the gesture can be considered "given"), the gesture is often produced smaller or in a less precise manner (Gerwing & Bavelas, 2004) and with a shorter duration (Holler et al., in press). It is important to note that all of the previously mentioned studies on gesture marking of referent status focus on specific types of referential gestures (or did not distinguish gesture type, e.g., Yoshioka, 2008). Indeed, Debreslioska & Gullberg (in press) mention the possibility that ISR may influence different subdimensions of gestures in different manners. As previously mentioned, non-referential (McNeill's "beat") gestures have been claimed to function as focus markers, in that they help visually mark focused or contrastive information in discourse (Kendon, 1980; Loehr, 2012; McNeill, 1992; Shattuck-Hufnagel et al., 2016).

The only empirical study to our knowledge investigating the non-referential gesture marking of ISR is by Im & Baumann (2020). Their study assessed the multimodal marking of ISR in a two-and-a-half-minute English TED Talk. In terms of ISR, the study used a more precise annotation of referents containing four levels: "new", "unused" (i.e., unique referents, such as "the Sun"), "bridging"

(which corresponds to accessible referents from context), and "given". Their descriptive analysis specifically focused on the relationships between gesture and pitch accent types, and gesture and ISR. Regarding the former, they found that L+H* pitch accents were most likely to co-occur with a non-referential gesture (59%), followed by H* pitch accents. The unaccented words were the least likely to receive a gesture (4%). Regarding the relationship with ISR, their results showed a tendency for non-referential gestures to mark more accessible ("bridging") and new ("new" + "unused") than given referents. The same database was analyzed for the relationship between prosody and ISR in Im et al. (2018). They found that pitch accents tend to be assigned to new and accessible referents, while given referents are generally unaccented. However, when assessing by pitch accent type, they found no significant relationship between pitch accent type and ISR categories. Taken together, these results suggest that while gesture associates with more prominent pitch accent types and newer information, pitch accent types themselves are not a reliable marker of ISR. As the authors did not analyze a three-way interaction between the variables (i.e., gesture, pitch accentuation, and ISR), the precise nature of this relationship remains unclear. As such, the main aim of **Chapter 5** of the current thesis will be to assess the joint marking of ISR via gesture and pitch accentuation in a corpus of English TED Talks. Importantly the coding of the database will control for prosodic features such as degrees of relative prosodic prominence, and will further assess the relationship between prenuclear pitch accentuation and gesture in the marking of ISR.

1.5. General objectives, research questions, and hypotheses

The main objective of the current thesis is two-fold. First, it proposes a novel approach to the study of co-speech gesture that espouses a dimensionalized view of gesture, where researchers should consider the semantic, pragmatic, and prosodic characteristics of gestures in an independent fashion and in a non-mutually exclusive manner. Second, it aims to further refine McNeill's phonological synchrony and pragmatic synchrony rules by better understanding the prosodic and pragmatic characteristics of both referential and non-referential gestures. To reach these two objectives, the body of this thesis is made up of four independent studies.

Regarding the first objective, the field of gesture studies has largely adopted McNeill's (1992) categorization of gestures composed of iconic, metaphoric, deictic, and beat gestures. Such a view classifies gestures in an unbalanced manner, characterizing the former three in terms of their referential and semantic properties, while characterizing the latter in terms of their relationship with speech prominence and pragmatic meaning. This is inherently at odds with the three synchrony rules initially laid out by the same author, where all gestures are semantically and pragmatically coherent with speech, and that all gestures are closely integrated with speech prosody. Following up on a set of recent proposals on the dimensionalization of gesture analysis (e.g., Prieto et al., 2018; Shattuck-Hufnagel & Prieto, 2019), the current thesis argues that

this approach is not ideal, particularly for corpus-based approaches to the study of gesture. Rather, each gesture should be assessed for each dimension independently, that is, its semantic contribution to speech, its pragmatic contribution to speech, and its association with speech prosody. By adopting this approach, gesture researchers will be in a better position to understand the complex relationships between co-speech gestures and speech. **Chapter 2** of this thesis will lay the foundation for the MultiModal MultiDimensional (M3D) approach to the annotation of co-speech gestures. Specifically, the chapter will justify the need for such an approach, assess a set of currently available labeling systems, and explain the specific aspects of M3D that will be followed throughout the subsequent chapters. The chapter will also introduce the two M3D-TED corpora which were developed using the M3D system and upon which the subsequent studies will be based.

The second main objective of the thesis will be assessed through a set of three empirical studies that will aim to further refine McNeill's phonological synchrony rule (in **Chapters 3 and 4**) and the pragmatic synchrony rule (in **Chapter 5**) A summary of the motivation, research questions, and hypotheses for each study are offered below.

The study in **Chapter 3** will assess the temporal association between gesture and pitch accentuation in French TED Talks by taking into account not only the role of the prosodic prominence in gesture attraction but also the role of prosodic phrasing, and specifically the AP domain. Importantly, most of the previous

studies have focused on English or other languages where pitch plays a prominence-lending role. Less is known about French, where pitch accentuation largely has a demarcative function, indicating the edges of the AP. Thus, the first aim of the study is to specifically assess two landmarks (the stroke and the apex) of all manual gesturals and their association with initial and final accents in natural (non-laboratory) French speech, comparing referential and non-referential gestures.

Moreover, much less is known about rhythmic patterns in the production of subsequent gestures. The only studies to our knowledge suggest that the production of rhythmic groups of subsequent gestures (henceforth RGGs) is largely independent from speech rhythm, with only one gesture within the group associating with the nuclear pitch accent (Loehr, 2007; McClave, 1994). The second aim of the study is to assess claims that RGGs composed of non-referential gestures are more rhythmic than RGGs composed of referential gestures, both in terms of their frequency, as well as their within-group isochronicity. Finally, the study assessed the relationship between speech rhythm and the production of RGGs. As such, we can identify the following three research questions:

1. Does pitch accentuation continue to act as a prosodic anchor for gesture in French, and is this relationship modulated by accent type (IA vs. FA) or gesture type (referential vs. non-referential)?

2. Do non-referential gestures have a tendency to form RGGs more than referential gestures, and are non-referential RGGs more isochronous than referential RGGs?
3. Do RGGs tend to mark subsequent APs in French, and is this relationship modulated by pitch accentuation (i.e., the presence or absence of IA)? If not, is this relationship sensitive to the temporal duration of prosodic phrases?

Regarding the first question, it is hypothesized that pitch accentuation will continue to act a prosodic anchor for gesture production, regardless of their demarcative function, showing similar tendencies to what has been described for English. It is believed that gestures will associate more with the IA in French, as these accents may serve more pragmatic or emphatic functions than FA, which mainly function to delimit the right edge of the AP. Given the tendency for alignment, it is hypothesized that the beat-like groups of gestures will also closely correspond to pitch accentuation, where there will generally be one gesture per AP (coinciding with FA), which may double when APs contain two pitch accents. Finally, we predict that beat-like groups of gestures will be equally made up of referential and non-referential gestures, showing no differences in isochrony between them as recent work has questioned the idea that certain gesture types are more closely related to rhythm and prosody (e.g., Shattuck-Hufnagel & Prieto, 2019). This study will thus contribute to the field by better understanding gesture-speech integration in a relatively understudied language in the field. It will contribute to our

understanding of the rhythmic, beat-like production of subsequent gestures, a subject that is largely neglected in the field.

Chapter 4 of this thesis will assess the temporal association between gesture and pitch accentuation in English TED Talks by taking into account phrasal prosodic structure. As we have noted in **subsection 1.3.2.4.**, early observations on the temporal association between gesture and pitch accentuation have suggested that gestures specifically associate with nuclear pitch accentuation (Kendon, 1980; McNeill, 1992). Thus, the study will specifically assess both the stroke and the apex gestural landmarks in natural English speech, comparing referential and non-referential gestures. The second aim of the study is to assess the role of nuclearity type (prenuclear vs. nuclear) in gesture-speech synchrony and whether this relationship is driven by relative prominence or phrasal positioning. Thus, we can identify the following three research questions:

1. Do gesture strokes and apexes align with pitch accented syllables in English TED Talks, and is this relationship modulated by referentiality?

2. Do gestures associate with nuclear pitch accents more than prenuclear pitch accents?

3. Is this relationship driven by relative prominence relationships or phrasal position?

Regarding the first research question, it is predicted that strokes will largely align with pitch accented syllables (as per Shattuck-Hufnagel & Ren, 2018), while apexes will show a smaller rate of alignment given that most studies on natural speech use rather broad criteria for alignment and Loehr (2004, 2012) reported a rather wide standard deviation in their temporal distribution. Regarding the second hypothesis, it is believed that prenuclear pitch accents will also serve as prosodic anchor points for gesture, as Loehr (2004, 2012) showed that G-Phrase onsets and ip onsets tend to temporally co-occur, suggesting that gesturing may occur early in the ip. Finally, we predict that these effects will not be driven by relative prominence, as prenuclear positions have been described as "attention-getting", marking the onset of a new prosodic phrase (e.g., Bolinger, 1985; Shattuck-Hufnagel et al., 1994), yet it is generally held that the nuclear pitch accent is structurally, and tends to also be phonetically, the most prominent in the phrase. This study contributes to the field by assessing how not only pitch accentuation, but also prosodic phrasing can be key factors in gesture-speech synchrony.

Finally, the study in **Chapter 5** will assess the multimodal cues to marking the ISR in English TED Talks. As we have seen in **subsection 1.4.**, most studies on the marking of ISR study one mode at a time (focusing on either prosody or gesture individually). Much work has evidenced the close relationship between prosody and gesture, thus the main aim of this study is to assess how gesture and prosody jointly work towards the marking of ISR. Moreover, most studies have focused on either the presence or absence of pitch

accents, or ToBI pitch accent type as markers of ISR. Another aim of this study is thus to assess the contributions of relative prominence as well as pitch accent type in the marking of ISR. In terms of the gestural marking of ISR, the study further aims to assess differences by referentiality, and to assess gestural marking in prenuclear positions. The following three research questions can be formulated:

1. How do gesture and pitch accentuation jointly mark ISR in English TED Talks?
2. In terms of prosody, what is the relationship between relative prominence, ToBI pitch accent type and ISR?
3. In terms of gesture, does gesture type (i.e., referential vs. non-referential) play a role in marking ISR? Are gestures sensitive to ISR status in prenuclear positions?

Regarding the first question, we hypothesize that pitch accentuation and gesture will largely go hand in hand to mark ISR, given their close relationship. In particular, new and (to a lesser degree) accessible referents will generally receive both types of multimodal cues. Given referents will be mostly unmarked by either cue. In terms of the second question, it is hypothesized that a probabilistic relationship will surface for pitch accent type, leading to a more stable relationship between ISR and relative prominence. Regarding the final research question, no significant differences between gesture referentiality types (i.e., referential and non-referential gestures) will be found in the marking of ISR. In prenuclear

positions, fewer new referents are expected to be found, causing gestures to associate more with accessible referents in prenuclear positions than given ones. The study will contribute to the field by disentangling the effects of relative prominence from pitch accent type for the marking of ISR, as well as elucidating the multimodal marking of ISR in prenuclear contexts, ultimately shining a light on the complex relationship between prosody, gesture, and the marking of ISR.

Finally, **Chapter 6** will discuss the main implications of the empirical studies in their respective fields of contribution, the theoretical advances offered by the M3D labeling system, as well as limitations and future directions of the present thesis. All in all, the three empirical studies in this thesis will not only contribute to our understanding of the prosodic association and pragmatic functions of gestures, but will also demonstrate how these prosodic and pragmatic characteristics are not relegated to a single gesture type. These empirical findings thus lend support to M3D as a valid approach to the study of gesture, where gestures are comprehensively assessed across three independent, non-mutually exclusive dimensions. Ultimately, M3D offers an opportunity to advance the field towards adopting a standardized, multidisciplinary approach to the study of gestures.

2

CHAPTER 2: MULTIDIMENSIONAL LABELING OF GESTURE IN COMMUNICATION — THE M3D PROPOSAL

2.1. Introduction

Speech is a multimodal act, where multiple modes of communication (e.g., verbal speech, speech prosody, gesture, etc.) are used as meaning-making strategies. These complementary vehicles of communicating meaning raise important questions about ways in which they participate in language. Addressing these questions requires extensive annotation of multiple modes and their interactions. Additionally, recent insights suggest that (manual) gestural behavior closely parallels prosodic structure and communicates both semantic and pragmatic meaning in a non-mutually exclusive manner. The present chapter introduces the MultiModal MultiDimensional (M3D) labeling system, which takes a novel approach to gesture annotation by accounting for three complementary and largely independent dimensions of gesture, which include Form properties, Prosodic properties, and Semantic/Pragmatic properties. The two main goals of this chapter are (a) to assess the features included in 10 currently available multimodal annotation systems; and (b) to describe the tripartite dimensional structure of M3D, a labeling system that is offered as an open access package that includes a set of reliable annotation conventions, training materials, and a 1-hour labeled audiovisual corpus. M3D is a tool that contributes to advancing the study of language as a multimodal phenomenon.

2.1.1. A holistic view of multimodality in language

Over its long history, the study of language has focused on the verbal aspect of speech. However, researchers today acknowledge that language is a *multimodal* phenomenon where multiple modes of communication (e.g., auditory and visual modes) come together to express meaning. Such a view has been supported by studies in the fields of sign language, neuroscience, language evolution, multilingualism, psycholinguistics, and development (see Perniss, 2018 for a review). The term "multimodality" specifically refers to all the different modalities (i.e., modes of communication) which are used as meaning-making strategies (e.g., Goodwin, 2000; Mondada, 2016). The field of gesture studies often uses the term multimodality to globally refer to both the verbal (or auditory) and visual modalities, entailing "what we hear" - oral speech - and "what we see" - co-speech gestures - as key meaning-making strategies in face-to-face communication. However, the terms "auditory" and "visual" are still rather broad categories and we can in fact identify multiple independent modes within these broad categories. For example, what we hear (the auditory mode) is in fact made up of two different modes: A speaker produces a morphosyntactic utterance (i.e., a string of words) which conveys meaning in its own right, and a second, superimposed layer of meaning that is added through the use of speech prosody. How these different meaning-making strategies independently contribute to communication can be seen when one single morphosyntactic string (e.g., the sentence "Dave is coming"), which can be produced using different intonational contours and co-speech gestures to

indicate different meanings to the listener. For example, in English it can be produced with a falling intonation to indicate that the utterance is a statement. The statement can be co-produced with a gesture to add supplemental information (for example, a "driving" gesture to indicate how Dave will be arriving). Alternatively, that same morphosyntactic string can be produced with a rising intonation to indicate a declarative question (e.g., Gunlogson, 2004). Again, that question could be produced with a gesture which adds further nuance to the meaning, such as a palm-up gesture to indicate the speaker is seeking a response, or a surprised facial expression to indicate the speaker's disbelief. Thus, the current study takes a holistic approach to multimodal communication, accounting for not only the verbal mode (i.e., morphosyntax), but also by including superimposed layers of meaning conferred via the mode of speech prosody (intonational patterns, tempo/rhythm, and intensity), as well as the visual mode (particularly manual gesture, facial expressions, and other meaningful bodily movements[3]). By using the term multimodality here, we want to emphasize the contribution of prosody and gesture in the creation of meaning in natural language as well as the interconnection between the two. And finally, we aim at underlining the relevance of simultaneously accounting for the multiple strategies when investigating meaning-making in language.

[3] In media studies, some researchers have used the visual mode to include the use of, e.g., images or Powerpoints, etc. While this is not incompatible with our vision of "multimodality" as meaning-making strategies, it goes beyond the scope of the current study.

If speech is truly multimodal, then as language researchers, we need to base our theories on a conception of language as an integrative multimodal phenomenon, and to understand the complex interactions that arise among different modes of communication. We need to assess ecologically valid contexts of multimodal language use and be able to annotate the different multimodal channels in a way that is shared and agreed across related subfields of linguistics (e.g., morphosyntax, pragmatics, prosodic and gesture studies, as well as discourse studies). However, these subfields are currently advancing independently, working separately from each other. As a result, there is a lack of speech annotation systems that account for shared aspects of form and meaning across all of these areas using conventional or widely-accepted standards. For example, the field of prosody has widely accepted the Autosegmental-Metrical (ToBI) system for prosodic annotation (a highly standardized procedure, with language-specific conventions for a number of typologically different languages; see Jun, 2005; 2014). While one objective in the field of prosody is to identify phonologically distinct intonational patterns and how they map to different pragmatic meanings (e.g., speech acts, focus marking, epistemic marking, etc.), ToBI does not necessarily espouse or recommend any specific pragmatic framework to annotate prosodic meanings such as epistemic stance categories like ignorance, certainty and uncertainty. Moreover, ToBI is a tool to annotate prosodic features of language, and prosodic labels in turn may be used to distinguish broad general pragmatic categories (e.g., declarative sentences, interrogation, vocatives, etc.). However, it

does not integrate more complex pragmatic meaning nor any sort of gesture analysis, despite the fact that research has shown a clear integration between, e.g., prosody and complex pragmatic meaning (such as stance), and between prosody and visual bodily features. This situation shows a need for more interdisciplinary and integrative approaches to multimodal speech annotation, as cross-disciplinary researchers would benefit greatly from having standardized annotation procedures (as offered by ToBI) integrated within a larger cross-domain annotation system.

In the last decades, the study of language and gesture (defined as "a visible action of any body part, when it is used as an utterance, or as part of an utterance," Kendon, 2004, p. 7) has been instrumental in widening our lens of investigation regarding the study of multimodal communication. Work by Kendon (1980) and McNeill (1992) was essential in establishing how speech and gesture are integrated temporally, semantically, and pragmatically. This close relationship has since been reinforced by studies across many sub-disciplines of linguistics. For example, numerous neurophysiological studies have found evidence that gestural cues are processed similarly to other aspects of language in terms of the semantic meaning (e.g., Özyürek et al., 2007 among others), and that information is treated similarly by the brain whether it is presented prosodically (via pitch accentuation) or visually (via gesture) (Biau et al., 2016). Further, the presence of such cues may in fact boost language processing (e.g., Weisberg et al., 2017), and developmental studies have shown how gesture production is often

predictive of later stages of language development (for an overview, see Hübscher & Prieto, 2019; Vilà-Giménez & Prieto, 2021). Such studies form the linguistic subfield of gesture studies, which aims to better understand the multimodal nature of language.

2.1.2. A multidimensional approach to gesture labeling

McNeill's (1992) classification system of gestures is perhaps the most widely accepted approach used by researchers in the field. It divides gestures into iconic, metaphoric, deictic, and beat gestures. Iconic gestures are those gestures which "bear a close formal relationship to the semantic content of speech" (p. 12) and are "pictorial" in nature (p. 14), illustrating concrete objects or events. Metaphoric gestures are also pictorial in nature, however, "the pictorial content presents an abstract idea … an image of the invisible … and image of an abstraction." (p. 14). Deictic gestures refer to pointing movements, which can indicate the location of objects or events in the immediate visual field as well as abstract pointing which may indicate abstract concepts in mental space. Finally, beat gestures do not portray or refer to any semantic content, but rather have been described as simple "flicks of the hand" up and down, or in and out seem to be moving with the rhythmic pulsation of speech and index words or phrases as being important for its discourse-pragmatic content" (p.15).

However, recent voices (e.g., Prieto et al., 2018; Shattuck-Hufnagel & Prieto, 2019) have acknowledged that the categorization of beat gestures is particularly problematic and non-accurate, as in addition

to describing its non-referential nature (i.e., that it does not portray or refer to semantic information in speech), McNeill describes beat gestures as having a particular hand configuration form, a one-to-one relationship with speech prosody, as well as special pragmatic functions. Importantly however, according to his own phonological and pragmatic synchrony rules, (a) all gestures are associated with prosodic prominence; and (b) any gesture type can convey pragmatic meaning. Regarding the former dimension, McNeill describes how all gestures occur just before or concurrently with the phonological peak syllable in speech. Regarding the latter dimension, the leading theories in gesture (i.e., Kendon, 2004; McNeill, 1992) convincingly show that the assessment of gestural meaning should include two independent aspects: semantic representation (i.e., referentiality, or how a gesture refers to semantic content in speech) and pragmatic meaning (i.e., how a gesture may signal information beyond merely semantic representation). Furthermore, the assessment of semantic meaning (henceforth referred to as gesture referentiality) and pragmatic function should be treated as non-mutually exclusive aspects of meaning, so that gestures can contribute multiple semantic and pragmatic meanings to an utterance (Lopez-Ozieblo, 2020; McNeill, 2006; Prieto et al., 2018; Shattuck-Hufnagel & Prieto, 2019).

In order to solve this categorization problem, Prieto et al. (2018) and Shattuck-Hufnagel & Prieto (2019) advocate for a multidimensional approach to gesture labeling that allows us to

disentangle the different properties of gesture. Such a dimensional approach suggests that gestures can be described in terms of three complementary yet largely independent dimensions, namely the form of a gesture (i.e., articulator, configuration, and kinematic properties), the prosodic properties of a gesture (gestural rhythmic, phrasing, and phasing properties, which are separate from an assessment of speech prosody or associations between the two modes), and gestures' contribution to meaning (both from a semantic and pragmatic perspective). Such a multidimensional approach is largely inspired by McNeill's clarification of his gesture classification system. In 2006, McNeill elaborates on his original gesture classification system, saying:

> I wish to claim, however, that none of these 'categories' is truly categorical. We should speak instead of *dimensions* and say iconicity, metaphoricity, deixis, 'temporal highlighting' (for beats), social interactivity, or some other equally unmellifluous (but accurate) terms conveying dimensionality.
>
> The essential clue that these are dimensions and not categories is that we often find iconicity, metaphoricity, deixis and other features mixing in the same gesture. Beats often combine with pointing, and many iconic gestures are also deictic. We cannot put them into a hierarchy without saying which categories are dominant, and in general this is impossible. A practical

> result of dimensionalizing is improvement in gesture coding, because it is no longer necessary to make forced decisions to fit each gesture occurrence into a single box. (p. 60, italics in original)

Following up on the proposals by Prieto et al. (2018) and Shattuck-Hufnagel & Prieto (2019), the position taken in the present study is that such a dimensional approach should not be limited to McNeill's original 4-dimension distinction which is limited to three semantic dimensions (iconicity, metaphoricity, and deixis) and one phonological one (temporal marking), but rather should apply more comprehensively, taking into account a gesture's prosodic, semantic, and pragmatic characteristics in an independent manner. Such a view puts McNeill's three synchrony rules at the forefront.

All in all, despite the fact that current research advocates for a truly holistic and multidimensional view of gesture analysis, first to our knowledge very few currently available gesture annotation systems have a multimodal view of data annotation that includes the morphosyntactic, the gesture, and the prosodic channels. Second, when focusing on gesture analysis, none of these systems provide a holistic and multidimensional analysis of gesture that includes its form properties, its prosodic properties (referring to aspects of gesture production that correlate with speech prosody), and its contribution to meaning.

2.1.3. Main features of the currently available gesture labeling systems

One of the goals of the present article will be to review a set of currently available multimodal annotation systems and assess their core features (see **Section 2.2.** of the current chapter for a review of 10 currently available annotation systems). Unlike the field of prosody which has largely (although not universally) come together around an established approach to annotation, the field of gesture studies has seen the development of many different annotation systems (see, e.g., Bressem, 2013; Ladewig & Bressem, 2013). Often, individual studies, projects, or research labs have developed their own labeling systems, and these systems vary in terms of which gestural features are coded and the approach used to coding them, making it challenging to compare data and results across studies. For example, we find that many core features of the articulator configuration and kinematic properties of hand movement (i.e., the physical description of movement and form, including aspects such as trajectory, handshape, position in space, etc.), as well as the segmentation of such movements into movement phases are shared and coded in a similar fashion in different systems, suggesting a certain degree of uniformity in coding for widely-agreed-upon form features (see **subsection 2.2.3.** below). Less than half of the systems (4 out of 10) mention the importance of including prosodic assessments, and systems differ substantially in how they approach the interpretation of gestural meaning (i.e., its semantic and pragmatic contributions in relation to speech). Furthermore, such approaches are often not in line with the

leading theories on (gesture) semantics/pragmatics. For example, only one system follow's McNeill's (1992) gesture classification, and about 60% of the reviewed labeling systems include some form of pragmatic labeling, though these are often very limited to specific research questions about particular functions (e.g., the role of gestures in turn-taking).

In our review of existing multimodal annotation systems, the features of accessibility, explicitness, and applicability of the multimodal annotation systems will also be systematically assessed. The first issue is 'accessibility', which specifically refers to how easy it is for researchers to access the description of the annotation system. Many annotation systems have often been conceived for use in a single funded project, so they may be presented on websites that are often not maintained beyond the duration of the project. Other systems may be described in stand-alone documentation (e.g., articles) which vary in their accessibility. Another issue is the 'explicitness' of the available descriptions, which refers to the amount of detail given in the descriptions. The documentation is generally a limited explanation of the structure of the system and the labels it uses. They often do not include clear step-by-step instructions or examples or tips for dealing with ambiguous situations (e.g., difficulty in identifying gestural movement phase boundaries). Finally, there is the issue of 'applicability', which refers to how easily a researcher can apply the system to a novel database. As many annotation systems have often been conceived for a single project, researchers looking to adopt a labeling system

are often forced to make adaptations in order to fit their needs. Rarely are there resources beyond the general description, so researchers may lack information necessary to apply the system to their own data (e.g., examples through annotated corpora, training materials, example post-annotation analysis, etc.). This pales in comparison to the field of prosody, where laboratories often share PRAAT scripts for data analyses, or have made online ToBI courses publicly available on stable websites which offer interactive training exercises for a variety of languages, including Mainstream American English (MAE-) ToBI (Veilleux et al., 2006), Cat_ToBI for Catalan (Aguilar et al., 2011) and P_ToBI for Portuguese (Frota et al., 2015), to name a few. Given these issues of accessibility, explicitness, and applicability, gesture researchers are often left to their own devices when considering how to approach gesture labeling in multimodal corpora.

All in all, to our knowledge, few previous multimodal annotation frameworks have adopted a holistic approach to multimodal corpora annotation which integrates the annotation of speech, gesture and prosody, and none has integrated a multidimensional conceptualization of gesture (including a multifunctional view in terms of semantics and pragmatics as non-mutually exclusive categories that contribute to meaning). Importantly, this view is based on the integration of theoretical frameworks of both McNeill (1992, 2006) and Kendon (2004, 2017) in that gestures can convey semantic meaning via iconicity, metaphoricity, and deixis, yet at the same time convey pragmatic meanings that are relevant in

communication, such as marking discourse structure, stance taking, or negation (see the **subsection 1.2.4.3**). Further, no system has aimed at being widely accessible; at being explicit by offering a wide range of detailed annotation procedures for a variety of central aspects of multimodal data coding such as prosody or pragmatics; and at being applicable to a variety of data by labelers who bring new research questions to the task.

2.1.4. Main goals

Given the current situation in the realm of multimodal annotation systems, the aim of the current chapter is twofold. The first goal is to review the main features of 10 currently available multimodal annotation systems, focusing on their goals, as well as accessibility, explicitness, and applicability issues. Such a review motivates the need for a more comprehensive and integrative approach to the annotation of multimodal corpora, which incorporates detailed annotation procedures and online training materials which are widely accessible to the community. The second goal is to present and describe the M3D (for MultiModal MuliDimensional) annotation system, which aims to address the need for a more comprehensive, integrative, and flexible approach to the annotation of multimodal corpora. This system has a number of features. First, it is comprehensive in that it espouses a holistic view of multimodality, explicitly calling for the independent but integrated annotation of speech, prosody, and gesture to highlight the complex relationships among them. Furthermore, M3D calls for a multidimensional view of gestures in a tripartite dimensional

structure (including form properties, prosodic properties, and gestural meaning). Namely, it allows for (a) the annotation of form properties such as hand configuration and kinematic properties for multiple bodily articulators, (b) the annotation of prosodic aspects of gesture, including prominence-lending movements, rhythmic movements, and phrasing properties, and (c) the multifunctional annotation of gestural meaning where gestures may convey both semantic and pragmatic meanings, and where gesture referentiality as well as their contextual pragmatic function can be annotated as non-mutually-exclusive categories. Second, it is integrative in that the system espouses the use of a number of standard annotating procedures for various aspects of speech, when available, allowing researchers from various fields to work in an interdisciplinary manner. Third, the system is flexible in that researchers may choose the different aspects of language that are relevant to label for their particular research agenda. Furthermore, the annotating procedures proposed vary in complexity, allowing for researchers to choose how complex such labeling should be. For example, researchers can annotate prosody using complex language-specific ToBI labeling, or less complex procedures such as Rapid Prosodic Transcription (Cole & Shattuck-Hufnagel, 2016). Finally, M3D represents an ongoing collaborative project that is currently housed on the Open Science Framework (OSF). It is publicly available and will include resources such as detailed annotation guidelines, an ELAN template, and a sample M3D-TED corpus of over 55 minutes of annotated speech for training purposes or as open data for scientific

investigation. Additional training material is currently in development.

The following subsection of this chapter will review ten multimodal annotation systems that have been developed in the field of gesture studies. The ten systems were chosen based on their relative ease of access, suggesting that these systems are readily available for researchers (further exclusion criteria are listed below). **Subsection 2.3** will then describe the main features of the M3D system and the M3D-TED corpus. Importantly, **subsection 2.3.6** will assess the reliability of some key parts of the system (namely, gesture phasing identification, apex placement, and crucially the non-mutually exclusive semantic and pragmatic labels).

2.2. Survey of existing multimodal annotation systems

The goal of the current section is to assess the main features of ten currently available annotation systems for multimodal communicative acts or interactions. The purpose is to assess the common and more standard features of multimodal annotation and the main differences between systems. The set of systems to be reviewed has been selected based on the following criteria:

1. The system has been thoroughly described in an accessible, standalone publication, whether that is a labeling manual, published article, or on a website (that is, the description of the system must go beyond the methods section of an

empirical study).

2. The system has been described so as to be applicable to a variety of multimodal corpora which can in turn be used by the research community at large.

3. The system covers at least one property of manual gestural annotation (e.g., form annotation, semantic/pragmatic meaning, etc.).

The section is organized as follows. First, a brief overview of the systems will lay out the various objectives for which each system was developed. Then, the aspects of gestural annotation for each system will be described, followed by an assessment of the different features of multimodal language behavior that are accounted for. Finally, an overview of the "Additional Characteristics" of each system (in terms of their accessibility, explicitness, and applicability) will be assessed. A set of visual tables have been added for each subsection to offer readers an easy-to-find reference guide which summarizes the presence or absence of various features across all systems.

2.2.1. Goals of the target multimodal annotation systems

Table 2.1 lists the annotation systems that will be assessed, along with their main goals, and the reference where a full description of the proposal can be found. All of the annotation systems were developed in the last two decades, are rooted in gesture and include

other aspects of speech to varying degrees. While some of the systems have very specific objectives, such as developing gestural taxonomies (CodGest), facilitating applications for automatic processing (CoGest, ASCG) or understanding multimodal pragmatic competence in older populations (CorpAGEst), other systems have very general goals, such as assessing gesture in terms of form and/or function (NEUROGES, MUMIN, LASG) or the interactions among different aspects of language (OTIM, DiaGest).

Annotation System	Reference	Goals
Outils de traitement d'information multimodale (OTIM)	*Blache et al. (2017).*	To better understand the interaction that exists between the different sources of information (i.e., prosody, lexicon, gesture, attitude, etc.).
Corpus of Academic Spoken English (CASE) annotation scheme	*Brunner & Diemer (2021)*	To develop a bottom-up manual annotation system for spoken language that takes into account previous research on multimodal features, focusing on salience and simplification and using a standard syntax.
The NEUROpsychological GESture coding system (NEUROGES®)	*Lausberg & Sloetjes (2009)*	To develop a tool for empirical gesture research that combines a kinetic with a functional analysis of gestural behavior.
CorpAGEst	*Bolly (2015)*	To establish the gestural and verbal profile of very old people in aging, looking at their pragmatic competence froma naturalistic perspective
MUMIN	*Allwood et al. (2007)*	To propose guidelines for a functional approach to gesture annotation in terms of feedback, turn-taking, and sequencing.
Annotation Scheme for Conversational Gestures (ASCG)	*Kipp et al. (2007)*	To aid in developing automatically generated and animated character-specific hand/arm gestures, making a conscious compromise between purely descriptive, high-resolution approaches and abstract interpretative approaches.
Conversational Gesture Transcription system (CoGesT)	*Trippel et al. (2004)*	To provide a transcription system for the linguistic analysis as well as automatic processing of conversational gestures
CodGest	*Maricchiolo et al. (2012)*	To develop a hand gesture taxonomy and a multi-media tool for coding in observational research.
DiaGest	*Jarmolowicz et al. (2007)*	To better understand interdependencies between gesture, lexicon, and prosody in Polish dialogues.
Linguistic Annotation System for Gestures (LASG)	*Bressem et al. (2013)*	To propose a systematic linguistic annotation for gestures grounded in a (cognitive) linguistic approach to language use and a form- based approach to gesture analysis that allows for an investigation of gestures' structures, meanings, and functions both on the level of gestures alone and in relation to speech.

Table 2.1: Goals of the assessed multimodal annotation systems

2.2.2. Coding of multimodal language: speech and speech prosody

Table 2.2 shows a graphical overview of the two features of speech that are included in the annotation systems. Speech coding refers to

the verbal mode. While it is assumed that any corpus analysis includes at least an orthographic transcription, NEUROGES, MUMIN, CodGest, and CoGesT make no mention of any sort of guidelines for transcribing verbal speech, and ASCG only transcribes the lexical affiliate of the gesture. CASE and CorpAGEst specifically include orthographic transcription guidelines (where CASE also accounts for paralinguistic features such as coughs, laughter, as well as camera changes and background movements). Other systems espouse more detailed analysis, such as transcribing syntax (OTIM, DiaGest), speech turns (LASG), and disfluencies (OTIM).

Annotation System	Speech coding	Prosodic Coding
OTIM	✓	✓
CASE	✓	~
NEUROGES®	✗	✗
CorpAGEst	✓	✗
MUMIN	✗	✗
ASCG	✓	✗
CoGesT	✗	✗
CodGest	✗	✗
DiaGest	✓	✓
LASG	✓	✓

Table 2.2: Overview of speech and prosodic coding included in the annotation systems. Green (✓) = feature is present, red (✗) = feature is not accounted for, yellow (~) = feature is mentioned but not explicitly accounted for.

In terms of prosodic coding, only three systems offer specific frameworks that should be followed: DiaGest proposes prosodic coding that accounts for major and minor intonational phrases (following Karpiński, 2006; Wagner, 2008, as cited in Karpiński et al., 2009), as well as strong and weak prosodic prominences (following the RaP system by Dilley & Brown, 2005). Prominences are then coded according to INSTINT (Hirst et al., 2000), a "sub-phonological", language-independent system that has been tested on a number of languages and can be executed automatically (Hirst, 2007). OTIM proposal to labeling French data is largely adapted from (Jun & Fougeron, 2005) to account for the most widely-agreed upon aspects of French prosodic phrasing, while prominence is also annotated according to INSTINT. LASG proposes the (obligatory) annotation of phrase or turn-final pitch movements and the (optional) annotation of focal pitch accents, both following the GAT2 conventions (Couper-Kuhlen & Barth-Weingarten, 2011). For its part, CASE mentions the importance of including prosodic information such as intonation, pitch, volume, speed, pauses, yet does not offer specific guidelines and merely includes such information as part of the speech transcription. None of the other systems (NEUROGES, CorpAGEst, MUMIN, ASCG, CoGesT, CodGest) include speech prosody. Summarizing, only three systems contain a fully multimodal approach to multimodal corpus annotation that includes the three channels, namely morphosyntax, prosody, and gesture.

2.2.3. Summary of coverage of gesture features across multimodal annotation systems

2.2.3.1. Gesture form, gesture phrasing and gesture phasing

Table 3 graphically summarizes the coverage of coding features for gestures. In each table, green (✓) indicates that the feature is present and specified to some detail, while red (✗) indicates the feature is not accounted for. Yellow (~) indicates that the feature is mentioned but not explicitly accounted for or specified in detail (all yellow cases are described specifically in the text). The coding of multiple articulators refers to whether systems account for coding different parts of the body (e.g., hand movements, head movements, eyebrow movements, facial expressions, body leans, etc.). The kinematic description refers to descriptions of the physical movement of the articulator in space. Aspects of gesture phasing and gesture phrasing refer to how movements can be grouped on various levels. Namely, gesture phasing refers to the annotation of component movements of gestures (e.g., gesture phases such as preparation, stroke, hold, etc.). Gesture phrasing, on the other hand, refers to how the annotation system describes coding for how gestures combine into larger constituents (e.g., Gesture Units; see Kendon, 1980; **Subsection 1.2.3.** of the current thesis, **subsection 2.3.3.3.** below). Semantic coding refers to the independent assessment of the semantic contribution of gestures (gesture referentiality) and pragmatic coding refers to the independent assessment of the pragmatic contribution of gestures (see **subsection 2.1.2.**).

Annotation System	Gestural Coding						Other Coding	
	Gesture Form		Prosodic Characteristics		Gesture Meaning			
	Multiple Articulators	Gesture Kinematics	Gesture Phasing	Gesture Phrasing	Semantic coding	Pragmatic coding	Speech coding	Prosodic Coding
OTIM	✓	✓	✓	✗	✓	✓	✓	✓
CASE	✓	~	✗	✗	✗	~	✓	~
NEUROGES®	✗	~	✗	✗	~	✓	✗	✗
CorpAGEst	✓	✓	✓	✗	✗	✓	✓	✗
MUMIN	✓	✓	✗	✗	✓	✓	✗	✗
ASCG	~	✓	✓	✓	~	✗	✓	✗
CoGesT	~	✓	✓	✗	✗	✗	✗	✗
CodGest	✗	~	✗	✗	~	~	✗	✗
DiaGest	✓	✓	✗	✗	✓	✓	✓	✓
LASG	✗	✓	✓	✓	~	✓	✓	✓

Table 2.3: Overview of the gestural features of the different gesture annotation systems. Green (✓) = feature is present, red (✗) = feature is not accounted for, yellow (~) = feature is mentioned but not explicitly accounted for.

The summary presented in **Table 2.3** shows that first most of the annotation systems acknowledge the importance of coding multiple articulators beyond the hands. While some researchers have shown interest in head movements in parallel with manual gestures, much less research accounts for other bodily movements that may participate in the communicative act (e.g., Shattuck-Hufnagel et al., 2010). To this end, OTIM, CASE, CorpAGEst, MUMIN, and DiaGest offer descriptions for the annotation of multiple articulators including (but not limited to) the hands, head movements, and posture. Both CoGesT and the ASCG acknowledge the importance of non-manual articulators, but reserve such details for future elaboration; meanwhile, NEUROGES, CodGest, and LASG only offer coding guidelines for the manual articulators.

In terms of coding the kinematic form (i.e., the physical description of movement and form, including aspects such as trajectory, handshape, position in space, etc.), seven of the systems have come together around a set of basic core principles, sharing common aspects and definitions such as handedness, handshape/orientation, and movement trajectory. The only difference between these systems in green is the amount of detail which is coded. For example, ASCG covers seven aspects of kinematic form (Handedness, Trajectory, Height, Distance, Radial orientation, arm swivel, hand-to-hand distance) while MUMIN only covers two aspects (Handedness and Trajectory). For their part, CASE, CodGest, and NEUROGES offer the most broad and/or least conventional labels. For example, CASE has approximately 40 form labels that include "throw away", "points to open hand", and may sometimes overlap with gesture referentiality (e.g., "beat", and the "peace sign" emblem, see below, as well as **section 1.2.2.** of the current thesis). Similarly, CodGest describes certain gestures in terms of a verbal description of their kinematic form (e.g., "pincers," "weaving," "whirlpool"). NEUROGES, for its part, does not code gesture kinematics per se, but rather instructs labelers to take a number of gesture kinematic features into account when determining the functional label for a gesture.

Aspects of gestural phasing are explicitly mentioned and coded in about half of the annotation systems. Specifically, six systems (OTIM, CorpAGEst, ASCG, CoGest, and LASG) account for movement phasing, largely following Kendon (2004), and Kita et al. (1997), which includes preparation, stroke, recovery/retraction,

and hold phases. Within these systems, only ASCG and LASG account for Kendon's higher level groupings (namely Gesticular Phrases and/or Units, see Kendon, 1980; **subsection 1.2.3.** of the current thesis). On the other hand, CoGesT only annotates the begin- and end-times of the gesture stroke. CASE, NEUROGES, MUMIN, CodGest, and DiaGest do not specifically mention gestural phasing.

2.2.3.2. Gesture meaning: Taxonomies, semantics and pragmatics

In terms of gesture meaning, the 'Semantic coding' column refers to the assessment of the semantic meaning in gestures and 'Pragmatic coding' column refers to the assessment of the pragmatic meaning of gestures (see **Section 2.1.2**). As previously mentioned, many systems make use of gesture meaning (as well as kinematics) in order to create a gesture classification system or taxonomy. Thus, the following subsections will first assess systems which independently annotate the semantic meanings of gestures, followed by those that assess the pragmatic meanings of gesture. Finally, a subsection will describe system-specific taxonomies, with a particular focus on which aspects form the basis of such classification.

2.2.3.2.1 Semantic coding

Three systems explicitly assess the semantic meanings of gestures. Surprisingly, only OTIM directly follows McNeill's (1992) typology of semantic meaning (i.e., iconic, metaphoric, deictic, and

beat gestures) though CodGest adopts this typology within their taxonomy for ideational gestures. MUMIN and DiaGest are based on Pierce's (1931) semiotic types: Indexical (deictic), Indexical (non-deictic, which would largely correspond to McNeill' "beat" gestures), Iconic, and Symbolic. However, these categories correspond quite closely to McNeill's Deictic, Beat, Iconic (and Metaphoric), and Emblematic gestures, respectively. Instead of assessing the semantic contributions of gesture, LASG assesses the relationship between semantic information in gesture and in speech in three ways. The 'temporal relation' assesses whether the gesture occurs before, simultaneously with, or after corresponding semantic information in speech, or if it occurs without speech. The 'semantic relation' determines whether information in gesture is redundant, complementary (or supplementary), contrary to, or replacing information in speech. The semantic relation annotation then determines a 'semantic function,' where, redundant gestures emphasize, complementary/supplementary gestures modify, contradictory gestures add, and replacing gestures substitute semantic information in speech. LASG also accounts for closely related aspects, such as modes of representation (acting vs. representing) and the identification of image schemas or motor patterns (though LASG describes this as an aspect of gesture form).

2.2.3.2.2 Pragmatic coding

A handful of systems include some pragmatic annotation of gesture, though they vary widely in terms of pragmatic areas and theoretical approaches, often related to specific research questions from the

developers. For example, MUMIN focuses on gestures that play a role in three specific pragmatic areas: turn-taking, feedback, and sequencing. As such, only gestures fulfilling these functions are annotated, and a set of additional sub-functional labels are then assigned to each gesture. For its part, OTIM includes annotations for discourse units and backchannels. CorpAGEst initially includes affective stance (i.e., emotion) annotation for facial expressions, though later publications have described the inclusion of more thorough pragmatic labeling to include discourse structure and stance-taking (Bolly & Boutet, 2018; Duboisdindien, 2019). In addition to annotating gestural turn-taking, LASG includes an assessment of speech act marking, particularly if the gesture is expressing propositional content, relates to illocutionary force, or affects the perlocutionary force of the utterance.

2.2.3.2.3. Gesture taxonomies

Four systems propose their own taxonomy for classifying gesture, which often combine aspects of semantic representation, pragmatic function, and kinematic characteristics. On the one hand, CASE and ASCG propose taxonomies that are largely based on kinematic form. CASE proposes approximately 40 descriptions described in general terms (e.g., "throw away", "points to", "open hand"), sometimes overlapping or coinciding with more typical classification schemes, such as McNeill's emblems ("peace sign") or beats. Similarly, ASCG proposes a list of 35 "lexemes" that typologize gesture, for example a "calm" gesture refers to gently pressing downward with palms facing downward. On the other

hand, NEUROGES and CodGest incorporate both form and meaning in the development of their taxonomies. NEUROGES proposes a functional classification of 28 different types of gesture. Labelers first determine the "function" of a gesture (which includes not only semantic representations such as indicating location or depicting object/motion, but also more pragmatic functions such as conveying emotion, organizing discourse structure, etc.). Once function is determined, labelers choose specific gesture types associated with that function (for example, gestures with a pointing function can be labeled as "deictic", "self-deictic", "body-deictic", "hand-showing" or "direction", and gestures functioning to add emphasis could be labeled as "batons", "back-toss", or "palm-out"). However, more detailed information about each subtype is not described. Similarly, CodGest proposes its own taxonomy where gestures can be divided into three main categories (i.e., cohesive, ideational, rhythmic). Cohesive subtypes are labeled based on their form (e.g., "weaving", "pincers"), while ideational subtypes relate back to typical McNeilian semantic subtypes of iconics, deictics, metaphorics, and emblems. Rhythmic gestures do not have any subtypes.

2.2.4. Accessibility, explicitness and applicability

Table 2.4 summarizes the additional characteristics in terms of accessibility, explicitness, and applicability. Regarding accessibility, eight out of the ten coding manuals/publications have been made available via open access, either through online repositories or academic social media. Only LASG is described as a

book chapter which has not been made available in open access, and the CodGest manual is uniquely available by directly contacting the authors. Five coding systems have their own webpages which offer a description of the project and a list of related publications (as well as the manual for download). However, most websites have not been updated or maintained since the project or grant has ended. For example, the OTIM and DiaGest websites were last updated in 2012, CorpAGEst in 2017, and CASE in 2018 when the projects ended. Though not necessarily problematic, this suggests that the development of these coding systems has largely stopped. The only regularly maintained website is that of NEUROGES, which gives details on how to access paid certification courses.

Annotation System	Accessibility		Explicitness			Applicability		
	Description available in open access?	Maintained on a website?	Description includes easy-to-follow, step-by-step guidelines?	Description includes examples of coding?	Publicly available annotated corpus?	Training material available?	Tips for post-annotation data management?	Reliability measures for the system?
OTIM	✓	✓	✗	~	✓	✗	~	✓
CASE	✓	✓	✗	✓	~	✗	✗	✓
NEUROGES®	✓	✓	✓	✗	✗	~	✗	✓
CorpAGEst	✓	✓	~	~	✓	✗	✗	~
MUMIN	✓	✗	~	~	✗	✗	✗	✓
ASCG	✓	✗	~	~	✗	✗	✗	✓
CoGesT	✓	✗	✗	✓	✗	✗	✗	~
CodGest	~	✗	✗	~	✗	✗	✗	✓
DiaGest	✓	✓	✗	✗	✗	✗	✗	✗
LASG	✗	✗	~	✗	✗	✗	✗	✗

Table 2.4: An overview of additional characteristics of the assessed annotation systems. Green (✓) = feature is present, red (✗) = feature is not accounted for, yellow (~) = feature is mentioned but not explicitly accounted for.

In terms of explicitness, first, only the NEUROGES system offers clear, step-by-step details on how annotators should label multimodal corpora, making use of question/answer flow charts for

coders to arrive at the correct label. While the CorpAGEst and CodGest systems do not offer step-by-step instructions, the manuals offer operational definitions, tips, and descriptions on how to code certain aspects (e.g., how to distinguish strokes, minimum durations to divide movement phases, etc.). For their part, LASG and ASCG describe the annotation system in the order in which labelers are to annotate (but does not offer specific annotation guidelines) and MUMIN offers a short description of "coding passes" that labelers may follow. Four of the annotation guides (i.e., OTIM, CASE, CoGesT, DiaGest) are much more general or theoretical in nature, providing a general framework, but without a set of instructions labelers should follow. Next, most of the coding systems offer examples, generally in the form of still images in the manual, although NEUROGES, LASG, and DiaGest do not offer any concrete examples. CASE and CodGest are the only systems which offer video examples. The former includes a snippet of the transcription so that readers can easily link the transcription to what is occurring in the video, while the latter uses video examples to illustrate each "gesture type" in their taxonomy. CodGest includes three video examples for each gesture type: an "ideal" example (filmed by actors), a "prototypical" (i.e., clear) example, and a "problematic" (i.e., dubious) example, the latter two coming directly from a corpus of news speech. Finally, while many of these systems have been applied to corpora, for only three systems are these corpora openly available (the CID corpus from OTIM and the CorpAGEst corpus are openly accessible for download on the

Ortolangue platform; CASE's VIMELF corpus is available upon request to the authors).

Finally, in terms of applicability, there are very few resources (apart from the descriptive manual) for novel researchers to better understand and apply the system to novel data. Systems that have been applied to open access corpora (i.e., OTIM, CorpAGEst, CASE) have descriptions regarding how to access the corpus from their websites, and only the OTIM website offers additional tools that may be helpful for researchers (such as automatic syllabifiers or phonetic transcribers, etc.). None of these systems offer any additional educational resources or tutorials on how to actually apply the labeling scheme to novel corpora, apart from paid personal or small group training seminars offered by NEUROGES® on their website. Furthermore, none of the systems offer any tips on how to manage data in post-annotation stages. OTIM (and to a lesser extent, DiaGest) comments on the use of XML format for interoperability purposes, but does not offer specific detail about how to manage the data from the various programs into XML format, or how to prepare data for any sort of analysis. Finally, the following six annotation systems offer reliability measures. OTIM found substantial reliability for prosodic annotations and automatic syntactic annotations, yet low rates of reliability for gesture space annotations (i.e., an aspect of kinematic form that encodes the location of gesture production relative to the speaker). MUMIN offers reliability measures for the annotations of a number of multimodal and communicative features related to turn-taking,

feedback, and sequencing (finding high rates of agreement for nearly all aspects except the annotation of head movements). NEUROGES and ASCG report substantial to high rates of agreement for gesture segmentation and classification, and CodGest also showed high rates of agreement for gesture classification. CASE also found high rates of reliability in their annotations, without specifying sub-features. CoGest mentions carrying out reliability analyses but does not describe any results, and CorpAGEst mentions the need to eventually assess reliability, as the coding has only been carried out by one labeler, but offers suggestions to ensure reliability in future work. LASG and DiaGest do not offer any evaluation of intercoder reliability.

In sum, the results of the present review show how the gesture research community has not developed a gesture annotation system that is widely agreed upon, specifically highlighting how the assessment of gestural meaning varies widely across the systems. Importantly, none of the systems reviewed follow the modern view that gestures should not be seen as pertaining to mutually-exclusive categories (e.g., McNeill, 2006). Additionally, currently available annotation systems do not systematically integrate the potential superimposed dimensions of gesture analysis (e.g., form, prosodic characteristics, and semantic/pragmatic meanings). While all of the systems assess multimodal language by including at least two modes of communication (speech and gesture), only three explicitly espouse a multimodal approach which aims to understand interactions between modes of communication. Even though OTIM

(and to a lesser extent, DiaGest) account for multiple visual and speech modes which come together to create meaning in an independent fashion, these systems still lack a thorough assessment of the potential pragmatic contributions of gesture, focusing more on interactions with syntax, prosody, or specific aspects of discourse.

Importantly, the community needs a system that reconciles several views into a more widely agreed gesture classification system and that is openly accessible, explicitly described, and easy to learn and apply to novel data. In the next section we will explain the motivation behind the current M3D labeling proposal, as well as the proposal itself, which aims at integrating our current knowledge on gesture analysis, their relationship with prosody, and their semantic and pragmatic contributions to speech. M3D aims to build upon the currently available systems by explicitly proposing a tripartite dimensional system to assess gestural characteristics in a largely independent, non-mutually exclusive manner. Furthermore, it integrates more standard and thorough assessments of the pragmatic contributions of gesture, and offers resources for its application to a range of databases. We believe that any multimodal annotation system should (1) adopt an approach which integrates flexible but standardized practices from multiple linguistic subfields to the annotation of multimodal corpora, (2) understand gesture as a set of multidimensional features that involve form, prosodic, and semantic and pragmatic features, and (3) be widely accessible, thorough, and offer support for its application to new corpora by novel researchers

studying multimodality. In this way researchers can develop comparable corpora that avoid the oversimplification of complex multimodal acts of communication, as well as opaqueness of their methods to other interested researchers.

2.3. Main features of the M3D annotation system

2.3.1. The concept of multimodal labeling

As previously mentioned, M3D is based on the well-supported idea that speech is multimodal and that three different modes (i.e., morphosyntax, prosody and gesture) interact with each other to create meaning for interlocutors.

Because such a multidimensional approach is rather interdisciplinary, it must remain flexible and adaptable to the individual researcher's goals. To this end, the annotation manual includes recommendations for many different levels of analysis, so that annotators may choose the specific levels they will focus on, depending on their research questions. Furthermore, one of the main advantages of M3D is that it incorporates several widely-accepted annotation methods that have been developed separately, uniting them into a single assessment tool. These include the use of widely accepted terminology, as well as standard annotation procedures (e.g. the independent assessment of different aspects of communication so as to avoid circular reasoning).

2.3.2. Multidimensional view of gesture

M3D incorporates the largely independent transcription of multiple modes of communication in multimodal language (namely, speech, prosody and gesture), and crucially it develops a multidimensional gesture annotation system. The three synchrony rules established by McNeill (1992) claim that gestural behaviors are integrated with prosodic prominence and communicate both semantic and pragmatic meaning, which highlights the importance of developing an approach that accounts for these multiple dimensions of speech-related gesture, including its kinematic form, prosodic properties, and semantic and pragmatic contributions. As such, the novel contribution of the M3D system is that it is grounded in three dimensions (as shown in **Figure 2.1**). These include 1) The *Gesture Form dimension*, which refers to a number of physical aspects of gesture across multiple articulators, including both configuration as well as the kinematic features of gesture, which include descriptions of the movement shape, direction, etc.; 2) The *Prosodic dimension*, which refers to the association of gestures with prosodic structure via a set of standardized perceptual procedures, prosodic characteristics of gestures, including beat-like-ness, phasing properties of gesture (the hierarchical organization of manual gestural movements) and 3) The *Meaning dimension*, which captures the semantic (i.e., referentiality) and pragmatic meanings that can be expressed in gesture. **Figure 2.1** shows a schematic representation of the three dimensions as well as their sub-features.

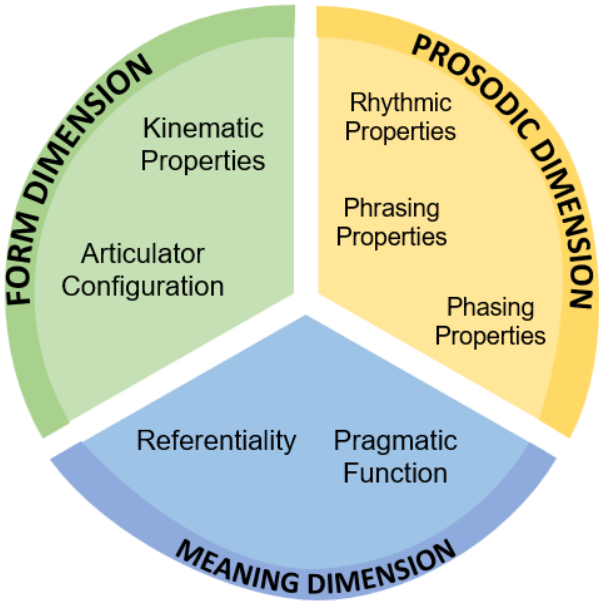

Figure 2.1: Overview of the M3D labeling system, including the main three dimensions and the properties included in each dimension.

In the following subsections, we briefly describe each dimension of the system, as well as the resources that are available to put the system to use, including a full labeling manual with specific procedures, ELAN[4] templates (Wittenberg et al., 2006), and additional training material. For more details about specific labeling guidelines, please refer to the full labeling manual available online[5].

[4] M3D has been developed for use in ELAN, but could potentially be adapted to any annotation software.
[5] https://osf.io/ankdx/

2.3.3. Gesture annotation

2.3.3.1. Identifying gesture

The perspective adopted here follows Kendon's definition of gesture as "a visible action of any body part, when it is used as an utterance, or as part of an utterance" (2004, p. 7). In other words, gestures are considered meaningful, communicative movements which can be produced by the hands, the head, facial expressions, or any other body part. Indeed, Kendon (1972, p. 204-205) discusses gesture as a full-body phenomenon. Examples of this include not only manual gestures which may convey semantic or pragmatic meaning. but also any communicative body movements, e.g. a body shift when used to indicate the addressee (see Sandler, 2018, p. 14), tilting the upper torso for pragmatic effect (Prieto et al., 2018; see also Shattuck-Hufnagel et al., 2010), or eyebrow movements to structure or emphasize information (Flecha-García, 2010; Swerts & Krahmer, 2010). Movements that are not intentionally meaningful (e.g., in some cases, scratching one's head, or adjusting one's clothes or hair) are not considered gestures and thus are not accounted for in M3D.

2.3.3.2. The form dimension of gesture

This dimension describes the physical nature of the gesture, in two main parts, whose labels are largely based on the *SCG Gesture Coding Manual*[6]. First, a tier set is dedicated to the *configuration of the articulators* that allows coders to indicate the predominant

[6] http://scg.mit.edu/gesture/coding-manual.html

gesturing hand, whether both hands are being used in a symmetrical manner, or even to code physical aspects of each hand separately. Subsequently, two tiers assess hand shape and palm orientation. A second tier set is dedicated to the *kinematic description of movement,* specifically assessing trajectory direction and shape. The above-mentioned tier sets are specific to manual articulators, however similar tier sets are available to code kinematic features of other articulators such as head movements, eyebrow movements, facial expressions, torso/body leans, etc. For example, a tier set for head movements includes annotations for trajectory direction and movement (Turn, Nod, Tilt, Protrusion, Slide, as per Wagner et al., 2014). **Figure 2.2** shows the annotation of the form dimension for a single non-referential gesture executed with the left hand. Specifically, the first tier shows the articulator that is being coded (here, LH for "left hand"). The next tier refers to the handshape, with the hand going from "fist" (F), to "relaxed" (R), and back to "fist" (F). The next tier is dedicated to annotating the palm orientation, with the palm starting orientated towards "self" (S), then "head" (H), then "up" (U), and ending at "self" (S). The next two tiers refer to kinematic properties, with the fourth tier showing the trajectory shape, and the fifth tier showing trajectory direction. Thus, the hand moves "straight" (S) "up" (U), then moves in a "curved" (C) shape down and forward (diag-DF), and ends moving straight (S) towards the "self" (S).

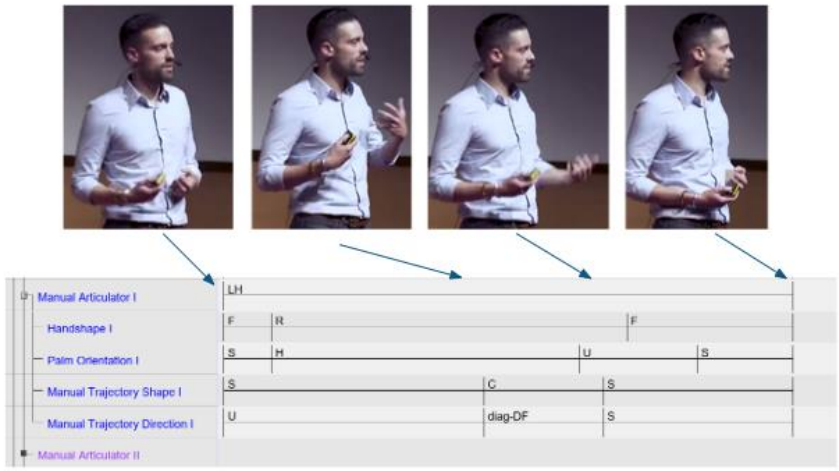

Figure 2.2: The annotation of the form dimension for a non-referential gesture (by taking the predominant left hand). Example taken from the French M3D-TED corpus, by speaker JP at (TEDx Talks, 2018) at 01:16.

2.3.3.3. The prosodic dimension of gesture

Gesture prominence and prosodic prominence have long been observed to be closely related (Kendon, 1980; Loehr, 2004; McNeill, 1992, among many others). One of the features of the prosodic dimension of gesture thus refers to the assessment of gesture movements which may be prominence-lending, as well as rhythmic properties of gesture and the phasing/phrasing properties of gesture. Importantly, these prosodic characteristics refer specifically to gesture, and assessments regarding speech prosody and any temporal association between the two are not accounted for specifically within this dimension. Importantly, M3D encourages independent assessments of speech prosody, which will be discussed in **subsection 2.3.3.4**.

First, in terms of prominence-lending and rhythmic properties, labelers may assess a gesture's "beat-like-ness" (e.g., Shattuck-Hufnagel & Ren, 2018). It refers to whether a gestural stroke is produced in such a manner that it is perceived as "accentuating", "punctuating", or marking a "rhythmic beat" - in other words, the movement phase of the stroke could be seen as having a prominence-lending function[7]. Such an assessment allows annotators to capture phenomena such as "superimposed beats", which have been described as beat-like movements that have been combined with (superimposed on) referential gestures (McNeill, 1992: 170). The evaluation of beat-like-ness is done without audio and is largely based on kinematic form, where each stroke is individually evaluated and labeled categorically as "very beat-like", "somewhat beat-like", or "not beat-like". Similarly, multiple subsequent strokes can seem to group together and mark rhythm in a beat-like fashion. These Rhythmic Groups of Gestures (RGGs) can be assessed independent of the referentiality of the gestures of which they are composed. For an example of an RGG, see **Figure 2.3** (see also **Chapter 3** of this thesis regarding the annotation of RGGs).

[7] Some aspects in the prosodic dimension may well overlap with the form dimension, being largely based on kinematic characteristics. However, such features seem to largely reflect relationships with speech prosody and for this reason we include them here.

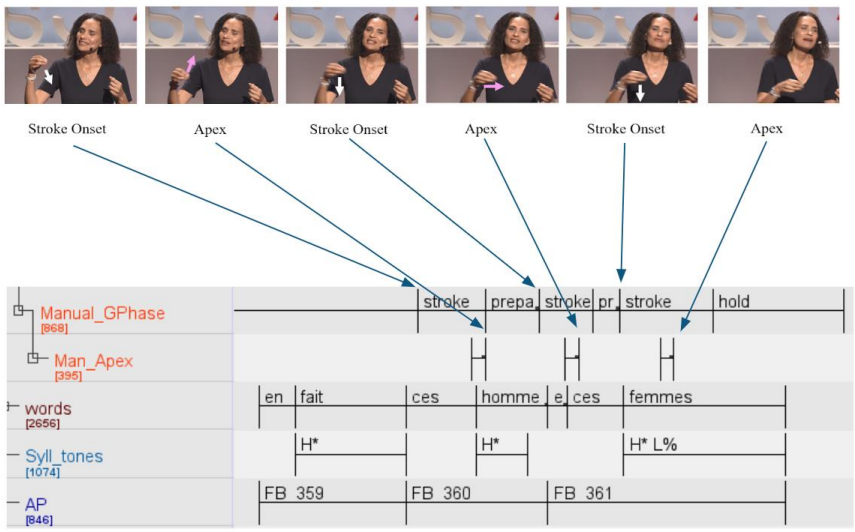

Figure 2.3: Example of an RGG produced with the utterance "En fait, ces hommes et ces femmes" ("In fact, these men and these women…") from the French M3D-TED corpus by speaker FB (TEDx Talks, 2015) at 06:54. **Upper panel**: Still-frames extracted at stroke onset and apex for each of the three gestures (arrows indicating direction of upcoming movement). **Lower panel**: ELAN annotations of the RGG, including gesture phasing, apex annotation, words. The latter two tiers refer to independent F-ToBI prosodic annotations for pitch accented syllables and prosodic phrases.

A second prosodic property of gesture refers to its phasing and phrasing properties. These two terms reflect the grouping of movements on various levels of higher-level structure. Specifically, phasing refers to the division of a single gesture into its component gesture phases (i.e., preparation, stroke, holds, and recovery). Alternatively, phrasing refers to the grouping of multiple subsequent gestures into larger units of movement (i.e., Gesture-Units). The grouping of movements into smaller or larger units has been shown to be organized in a parallel fashion with prosodic

phrasing structure, where each level in the prosodic hierarchy is associated with distinctive patterns of bodily movement in the gestural hierarchy (Kendon, 1980; see also Shattuck-Hufnagel & Ren, 2018). Thus, we include the description of gesture phasing as a prosodic aspect of gesture, though it is also relevant in the other dimensions (being largely derived from the form dimension, as well as the domain in which labelers assess the meaning dimension). Adopting standard terminology from Kendon (1980) and Kita et al. (1997), M3D calls for the labeling of the gesture unit (G-unit), which refers to the span of time from when the hands leave rest until their return to rest[8]. **Figure 2.2** shows an example of a G-unit, which includes two gesture strokes representing all communicative movements from when the hands leave a rest position, until their return.

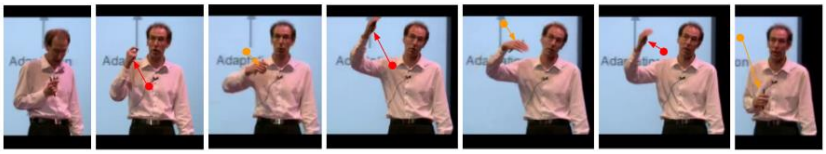

Figure 2.2: A full G-unit, taken from Keith (2007) at 03:34. Upward movements are indicated by red arrows, and downward movements are indicated by orange arrows.

Gesture units are then subdivided into the smaller level of phasing, where individual movements or gestural "movement phases" are

[8] "Rest" here refers to moments when the speaker is not actively moving the hands or maintaining them in an active position - rest positions may vary by speaker and across time within one speaker, and include having the hands down by the waist, held in front of the speaker, or in any other position where the speaker is deemed to be not actively in the process of gesturing.

classified as a preparation (movements from rest), stroke (the most prominent movement which bears communicative meaning), hold (pauses in movements, generally before or after the stroke; the only obligatory phase for a movement to be considered a gesture), or recovery (a return to rest). Finally, the apex of the stroke is also identified, which can be defined as the kinetic goal of the stroke (Loehr, 2007, p. 189) and can be seen as points of maximum extension, sudden stops, or changes in direction (see also Yassinik et al., 2004). **Figure 2.3** shows an example of multiple gesture phases (upper panel) as well as the apex (lower panel) of a single gesture. Literature in the field of gesture studies has described the level of grouping called the *gesticular phrase* which includes the stroke and any potential preparation or hold associated with that stroke. However, this level of phasing can be recovered automatically from gesture phase labels outside of ELAN and thus are not coded in M3D.

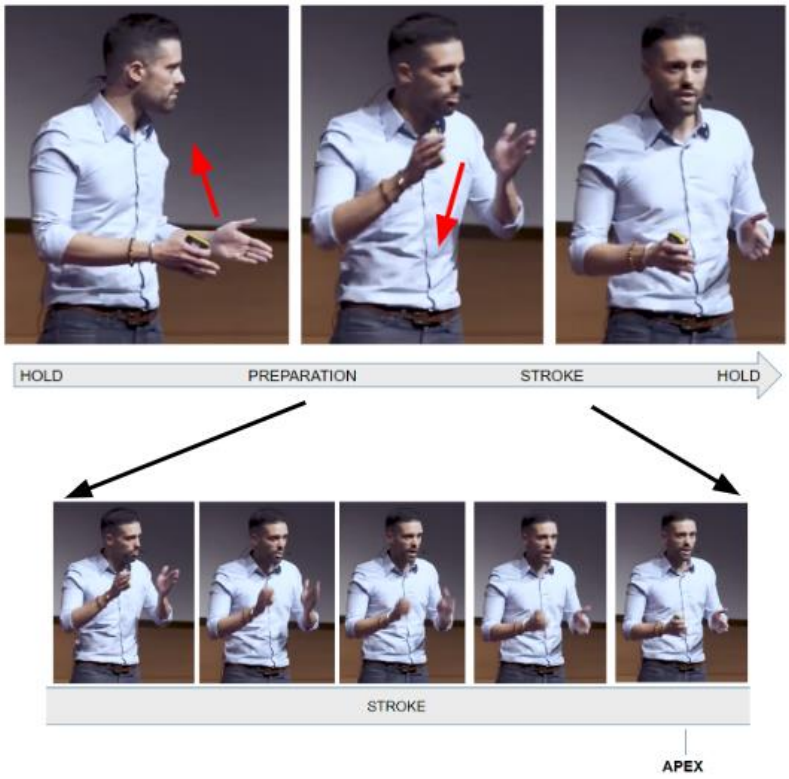

Figure 2.3: Still images of a gesture executed in the French M3D-TED corpus, by speaker JP (TEDx Talks, 2018) at 02:58. **Upper panel:** the various gesture phases involved in the execution of the gesture. **Lower panel:** frame-by-frame images of the stroke, where the final frame indicates the apex.

2.3.3.4. Independent analyses of speech prosody and association between the two modes

A holistic evaluation of each gesture might want to assess whether it is perceived to be associated with prominence in speech or not. This perceptual evaluation of the link between the meaningful gestural movements and speech prominence has been termed the "Prominence Association Component" (PAC), which allows for a

loose assessment of timing and meaning associations between gesture and speech prosody (for data comparing the two measures, see Rohrer et al., 2019). The annotation of the PAC involves watching a gesture while listening to the concomitant speech, and judging if the gestural movements align with prosodically prominent syllables. Such an approach may be more suitable for labelers who lack any sort of prosodic annotation training.

Previous studies have also shown how gestures can influence the perception of speech prominence (Krahmer & Swerts, 2007), and that the perception of gesture-speech alignment by human evaluators is not very reliable (Leonard & Cummins, 2011). For these reasons, M3D recommends that annotations for speech prosody be carried out independently from video (that is, outside of ELAN so as to not have access to video data, e.g., in Praat; Boersma & Weenink, 2022). This allows for precise quantitative analyses of the acoustic information (e.g., distance in ms to nearest landmark, etc.). While researchers may choose any sort of prosodic labeling system (including automatic labeling systems such as INTSINT, see Hirst, 2007), M3D recommends two systems in particular for manual annotation that researchers can choose from. Researchers may use full ToBI systems (with Tones and Break Indices, Silverman et al., 1992; see also Jun, 2005, 2014) to identify tonal targets associated with phrasally-stressed syllables, as well as with the boundaries of a set of hierarchical intonational constituents. Another option is Rapid Prosodic Transcription (RPT, Cole & Shattuck-Hufnagel, 2016), a somewhat less comprehensive system which makes use of “crowd-sourced” identifications of prominence

and boundaries. This technique is highly reliable for identifying prominences and phrasing; however, it does not offer such details as precise tonal targets. For example, RPT (or other manually-corrected automatic systems) would be a good option for researchers who are not specialists in prosodic annotations, yet who would like to include identification of prosodic patterns in their analyses.

2.3.3.5. The meaning dimension of gesture

The meaning dimension includes two main parts, namely (a) the semantic contribution of gestures (in terms of their referentiality) and (b) their pragmatic functions, e.g. their role as markers of various pragmatic meanings such as stance, information and discourse structure, etc. (see **subsection 2.3.2.3.2.**). As previously mentioned in **Section 2.1.2.**, M3D's approach sees gestures as potentially contributing to both semantic and pragmatic meaning to speech in a non-mutually exclusive manner (that is, a single gesture may represent propositional content, which does not automatically preclude it from also contributing to pragmatic meaning, and vice versa). Such an approach aims to integrate the two theoretical frameworks laid out by McNeill (1992) and Kendon (2004). By incorporating the semantic (referentiality) and the pragmatic dimensions of gesture when describing their meaning, we are incorporating both views. First, by taking on board McNeill's use of referentiality properties we propose a basic divide between two types of gestures, that is, referential and non-referential gestures. Second, by incorporating a gesture pragmatics tier set we take on

board Kendon's view on how gestures may convey a variety of pragmatic functions in discourse.

2.3.3.5.1. Gesture referentiality

In terms of semantic meaning, M3D is largely based on McNeill (1992, 2006) where he distinguishes iconic, metaphoric, deictic, and beat gestures. An important novelty of M3D deals with our approach to the latter gesture type. While the former three are categorized based on their semantic (e.g., referentiality) properties, beat gestures have been defined as composites of three properties, namely in that (a) they do not represent any semantic content, (b) they have a pre-defined form (up-down or in-out flicks), and (c) they have a strong relationship with speech prosody (rhythmic movements). Given that the latter aspect has been shown to apply to all gesture types (see Prieto et al., 2018; Shattuck-Hufnagel & Prieto, 2019; Shattuck-Hufnagel & Ren, 2018) and that M3D captures these aspects in an independent manner (namely through the Form and Prosodic dimensions), M3D adopts a broader view in terms of typologizing gestures based on their semantic contribution, or lack thereof. We call this gesture referentiality, and we can distinguish two broad types of gesture. Referential gestures are those which integrate aspects of semantic meaning based on a clear referent in speech (e.g., see **Figure 2.4 below**). Non-referential gestures, on the other hand, refer to any gesture that does not have a clear and direct link to the semantic content in speech, crucially regardless of its gesture form or relationship with speech prosody

(e.g., see **Figure 2.3 above**). Emblems are also coded within this tierset.

Another novelty of M3D is that referential gestures can then be further assessed in terms of dimensions of iconicity, metaphoricity, and deixis as non-mutually exclusive categories (as per McNeill, 2006; see **subsection 2.1.2.**). For example, **Figure 2.4** shows an example gesture taken from series of five gestures demonstrating iconicity and metaphoricity, being performed in a very beat-like manner. The speaker says "Finally, we could make the particles migrate to **over** the **poles**, so we could arrange the climate engineering so it **really focused** on the **poles**" and performs a series of gestures (their positions indicated in **bold**). The gestures show degrees of iconicity, as it iconically represents poles sticking out at the ends of the Earth. The gesturing is also metaphoric in that it is representing an abstract concept, as the poles of the earth are not actual poles but rather refer to specific geographic/magnetic points on Earth. Finally, the speaker produces the gestures in a beat-like fashion, loosely marking speech rhythm. This approach is not implemented in any of the previously reviewed annotation systems (potentially due to the fact that codifying behavioral phenomena into mutually-exclusive categories could be considered beneficial for reliability purposes, see Maricchiolo et al., 2012; see also **subsection 2.3.6.**).

Figure 2.4: An example of a gesture taken from a series of five gestures showing iconicity, metaphoricity, and beat-like-ness, taken from Keith (2007) at 08:34.

In order to code gesture referentiality as dimensions in ELAN, each referential meaning has its own tier, which allows labelers to annotate any potential combination of the referential semantic meanings present in a superimposed manner. For researchers in the field of gesture studies who wish to continue working with a set typology of gestures (for example, to be able to count the number of "pointing gestures"), M3D provides this option, e.g. researchers may choose to count all of the gestures that show degrees of deixis, even though some of them also show additional referential dimensions such as metaphoricity (e.g., pointing upward to metaphorically represent an abstract increase). Alternatively, labelers may add an additional "predominant semantic" tier where in addition to assessing the various semantic dimensions, they annotate the dimension that seems the most predominant, or simply choose only one label when annotating referentiality. These

examples highlight that M3D is highly flexible and adaptable to the researcher's needs, and does not force researchers to adopt pre-established gesture typologies for gesture annotation.

2.3.3.5.2. Pragmatic functions of gestures

The literature on gesture has started to establish a wide range of pragmatic functions that gestures carry out. Kendon (2004, 2017) has qualitatively assessed the numerous pragmatic functions a gesture may fulfill, which led him to develop a set of pragmatic functions of gesture, which included aspects such as operational functions, modal functions, performative functions, and parsing functions. Other early researchers describe interactive conversational gestures which contribute to the "nature of dialogue itself, rather than with the specific topic of discourse" (Bavelas et al., 1992, p. 476). However, most studies describing the pragmatic functions of gestures often do so through the perspective of form, discussing "palm-up open hand" gestures (e.g., Cooperrider et al., 2018; Ferré, 2011), "hand flips," "finger bunch," or "ring" gestures (Kendon, 1995, 2004). McNeill (1992) also briefly mentions some discourse-pragmatic effects (mainly in relation to beat gestures), namely describing how they can mark "the introduction of new characters, summarizing the action, introducing new themes, etc." (p. 15).

The pragmatic functions of gestures proposed by M3D have been developed based on a review of the pertinent literature on gesture pragmatics (e.g., Bolly & Boutet, 2018; Brown & Prieto, 2021;

Kendon, 2004, 2017). As such, the pragmatic functions described are based on the most common functions that have been identified in the field of gesture and have been linked to standard sub-fields in pragmatics. By doing so, five main pragmatic domains have been identified and an initial subset of functions have been described (see **Table 2.5**).

Pragmatic Domain	**Pragmatic Functions**	**Sub-functions**
Discourse Organization	Information Structure	Focus/Topic/Information Status of Referents/Contrast
	Discourse Structure	Parenthetical, Listing, etc.
Operation	Affirmation/Negation	
Stance	Epistemic Stance	Certainty vs. Uncertainty
	Affective Stance	"Excited," "angry," etc.
	Politeness Stance	
	…	
Speech Act	Illocutionary/Perlocutionary	Speech act type
Interactional	Turn-taking	Turn holding, requesting, etc.

Table 2.5: The five pragmatic domains as well as selected functions and sub-functions that have been identified for M3D.

First, speakers can use gestures to mark how they organize their discourse (see also **Subsection 1.4.3** as well as **Chapter 5** for a review on the gestural marking of IS). The **Discourse Organization** domain thus relies on two subfields of pragmatics: Information structure (e.g., Krifka, 2008) and discourse structure (e.g., Grosz & Sidner, 1986). In terms of information structure, M3D proposes adopting the labeling scheme described in Götze et al. (2007), as it contains guidelines for both simple and complex annotations of information structure across three levels (Focus, Topic, Information status of referents), based on the text and independent of speech prosody. The labels within discourse structure marking have been collected from the relevant literature on the gestural marking of discourse structure and include relevant labels such as gesturally marking the start or end of sequences, parenthetical digression, listing, or using gesture as a cohesive device (e.g., Bolly & Boutet, 2018; Kendon, 2017; Ladewig, 2014).

Speakers may also use gestures to affirm or negate (e.g., Prieto & Espinal, 2020). Any sort of multimodal affirmation or negation would fall under the domain of **Operation**. Largely based on Kendon (2017), affirmation refers to any gesture that expresses a positive interpretation, while negation refers to any gesture that expresses a negative interpretation. This can either be fairly straight forward (a hand sweep while saying “There’s no more brie.”) or more abstract in nature (Saying “if it’s good quality stone, you’ll use it again because it’ll last forever”, while shaking the head

during “because it’ll last” -- this shaking implies the stone will *not* decay, it will *not* become unusable)[9].

Speakers may use gestures to show their stance in regards to their discourse. The **Stance** domain englobes a broad view of stance as “personal feelings, attitudes, value judgments, or assessments” (Biber et al., 1999, p. 966, as cited in Freeman, 2015). Du Bois (2007) proposes the “Stance Triangle”, describing how stance-taking is a three-part act where speakers (1) evaluate a stance object, (2) position a subject (usually oneself) in terms of that stance object, and (3) align stances between interlocutors. There are limitless ways in which stance can be expressed and in line with this broad view, Du Bois (2007) states:

> Because of the diversity of observable stances in principle without limit, it is necessary to go beyond merely cataloguing their contents or classifying their types. To frame a theory of stance means to provide a general account of the mode of production of any stance and of its interpretation in a context of interaction. (p. 192)

As such, the complex undertaking of listing all of the possible multimodal stance marking that can be carried out is not the objective for M3D. However, an initial subset of stance labels have

[9] Examples taken from Kendon (2017, p. 170) -- Keep in mind that this latter example also includes a stance-taking function (an evaluation of the stone’s quality).

been proposed to account for areas such as epistemic stance, affective stance, and politeness stance (the evaluating and positioning aspects of the stance triangle), as well as agreeing and cooperating (the alignment aspect of the stance triangle). For studies describing the multimodal expression of stance-taking, see, e.g., Bolly & Boutet (2018), Borràs-Comes et al. (2019), Brown & Prieto (2017), Brown & Prieto (2021), Cooperrider et al. (2018), Crespo-Sendra et al. (2013), Esteve-Gibert & Prieto (2018), Ferré (2011), Freigang & Kopp (2015, 2017), Hübscher et al. (2020), Kendon (2017), Ladewig (2014), Roseano et al. (2016), among many others.

Speakers may also use gesture to perform speech acts. The **Speech Act** domain, thus, identifies whether the gesture is related to the illocutionary force or affecting the perlocutionary force of the speech act (e.g., Bressem, 2013) as well as the type of speech act that is being produced (speech act types from Searle, 1975). Finally, the **Interactional** domain refers to when speakers use gestures to manage discourse in interaction with interlocutors, namely via turn-taking (e.g., Levinson & Torreira, 2015; Sacks et al., 1974).

Non-mutual exclusivity is also present within this dimension. Just as referential gestures can contain a mix of different degrees of iconicity, metaphoricity, or deixis, a single gesture can fulfill multiple pragmatic functions simultaneously (see, e.g., Lopez-Ozieblo, 2020 regarding the multi-pragmatic effects of gestures). For example, if a speaker says “We were very excited about this” and concurrently produces a non-referential gesture where both

arms move downwards and outwards occurring with the word "very", that gesture would be interpreted as showing both stance marking ("affective stance" in that the speaker is showing his excitement visually) as well as operational marking ("negation" in that the speaker is denying that any other thing could augment their excitement - for similar examples, see Kendon, 2017). Both meanings would be annotated as being carried out by the manual articulators as well as prosodic prominence, and the affective stance would be considered "strong" while the operational marking would be considered "weak".

It is also important to reiterate that M3D accounts for an assessment of the pragmatic meaning that is independent of the semantic meaning and the gesture form. Referential gestures may also contribute pragmatic meaning to speech, as Kendon (2017) indeed describes how the handshape of deictic gestures may change as a result of how the pointing is being used in discourse (for example, to distinguish two different objects vs. to comment on an object). Additionally, pragmatic functions are not limited to particular hand forms (e.g., "palm up open hand gestures," see above). This is evidenced in the often overlapping and even contradictory functions that form-based analyses have shown (for example, a palm up open hand form has been linked to both lacking knowledge as well as showing obviousness, e.g., Cooperrider et al., 2018).

All in all, there is a clear need for the gesture field to adopt an annotation system that accounts for independent assessment of both

the semantic as well as pragmatic contributions of gestures within widely-accepted theoretical approaches, allowing gesture researchers across the field to make use of the same tool which can be adapted for their own needs in their assessment of gestural communication.

2.3.4. M3D accessibility, explicitness, and applicability

In order to provide a stable, standard method of multimodal labeling, M3D has been designed to be easily accessible online, explicit and learnable. It offers explicit details to aid the user, including tips on what to do in cases of ambiguity during annotation, tips/tutorials dealing with how to work with the different tools to carry out annotation), and information regarding how to deal with data analysis post-annotation.

2.3.4.1 Accessibility

The M3D project is hosted as an open-access permanent platform on the OpenScience Framework (OSF)[10] under the title "TheMultiModal MultiDimensional (M3D) labeling system" and is regularly updated. **Table 2.6** lists the resources available on the OSF website.

[10] https://osf.io/ankdx/

Materials available online in the OSF page "TheMultiModal MultiDimensional (M3D) labeling system"	
Resource	**Comments**
M3D Labeling Manual	An updated version will be released together with the publication of this article. Includes step-by-step instructions, workflow tips, and examples.
M3D ELAN template	Includes Tier Hierarchies and Controlled Vocabularies. Researchers can start coding right away.

English M3D-TED corpus	Time-aligned dataset of TED Talks in American English (5 speakers, totaling over 23 minutes of multimodal speech) containing the following annotations:
	• Time-aligned orthographic transcription
	• ToBI prosodic annotations, PAC
	• Annotations regarding Information Structure (Information Status of Referents)
	• Gesture annotations:
	  ○ Gesture prosodic features: Gesture phrasing, phasing, apex, beat-like-ness
	  ○ Gesture meaning: Gesture Referentiality
	  ○ Gesture meaning: Pragmatic annotations regarding gestural marking of Information Structure (Information Status of Referents, contrast)

French M3D-TED corpus	Time-aligned dataset of TED Talks in Metropolitan French (5 speakers of each language, totaling over 37 minutes of multimodal speech) containing the following annotations: • Time-aligned orthographic transcription

	• ToBI prosodic annotations, PAC
	• Annotations regarding Information Structure (Information Status of Referents)
	• Gesture annotations:
	  o Gesture prosodic features: Gesture phrasing, phasing, apex, beat-like-ness, RGGs
	  o Gesture meaning: Gesture Referentiality

Training Materials	A series of GIFS, short tutorial videos, and practical exercises to train multiple aspects of M3D. Under development (initial publication date in late 2022)
OSF links to other sub-projects using M3D or M3D-TED corpus	Resources for post-annotation data management (in R)

Table 2.6: Resources available on the M3D OSF website.

Specifically, the website hosts the detailed M3D Labeling Manual, ELAN template, M3D-TED corpus, and additional training materials, as well as a number of sub-projects that host studies that have used M3D or the M3D corpus. All of this material is openly accessible and will be regularly updated for use by the research community.

2.3.4.2 Explicitness

The M3D labeling manual contains a detailed description of the system (along with theoretical justifications and a bibliography). It also offers detailed step-by-step instructions that can be adapted

according to each individual researcher's objectives. In addition to step-by-step procedures, many aspects of the system contain examples linked to actual real-world data for clarification for any novel coders, and cases of ambiguity that have come up when applying M3D to the M3D-TED corpus have been identified. A set of "tips" to overcome such difficulties have been included as in the manual. In addition to the manual, an ELAN template file is also available which contains tiers for every aspect of coding described in the M3D labeling manual, organized in hierarchical order (i.e., organized as parent/child tiers for coding each dimension, that is the form, prosodic, and meaning dimensions) and associated with all of the controlled vocabularies available. As such, researchers can download the template and begin working with M3D right away.

2.3.4.3 Applicability

In terms of applicability, the OSF page will also link to a set of training materials for novel researchers (currently under development). These materials will include video-recorded tutorials on M3D and its dimensions, labeling tutorials so that novel researchers can better understand the workflow with M3D, as well as GIFS and screenshots to understand how to work with different tools such as ELAN and Praat. An online platform with practical exercises will also be developed so that learners can practice annotating and receive automatic feedback. The OSF page also contains links to sub-project pages (i.e., "components") that host studies that have made use of M3D or the two M3D-TED corpora. By having access to these pages, novel researchers can see how

M3D has been applied in practice, as well as access, e.g., code scripts for data analyses. Such information may aid novel researchers in post-annotation data management.

The next section will describe the M3D-TED corpora, two rich multimodal, time-aligned corpora that have been annotated following M3D standards. These resources have been made available to the research community at large with multiple aims. First, they can be an additional resource for learners of M3D, as they offer a large amount of M3D coded material. Secondly, researchers may feel free to use the data to run their own studies, or create additional annotations

2.3.5. M3D-TED French and English Corpora

The M3D-TED corpora refers specifically to the individual English M3D-TED corpus and the French M3D-TED corpora, and are the result of applying M3D to real data, specifically TED Talks. Currently, the two M3D-TED corpora contain 10 adult speakers (five in American English and five in Metropolitan French), and represent approximately 61 minutes of multimodal speech. As this is a continually evolving project, approximately five minutes of multimodal speech has been coded per speaker. These stretches of speech were chosen based on a number of criteria, namely if the speaker was regularly visible due to changes in the camera angle, if they produced gestures, and in function of the total length of the talk. The set of annotations includes a time-aligned orthographic transcription, full ToBI annotations for speech prosody, and

annotations for Information Structure (particularly for the information status of referents and contrastive elements). Specifically for gesture, the corpora include annotations for gesture phrasing, phasing, apex, beat-like-ness, rhythmic groupings of gestures (RGGs), and gesture referentiality (see **Table 2.6**). Future plans for the M3D-TED corpus include the annotation of the form dimension, as well as more thorough coding of gesture pragmatics.

2.3.5.1. Description of the French and English M3D-TED corpus annotations

Table 2.7 shows the descriptive statistics of gesture production across the two individual corpora and speakers in terms of number of gestures, as well as gesture rate in terms of words per gesture and gestures per minute.

Speaker (Sex)	M3D_TED Corpus	Amount of multimodal speech (h:m:s)	N of Gesture Strokes	Words per Gesture Stroke	Gesture Stroke per Minute
AS (M)	English	00:04:02	163	4.61	40.45
EG (F)	English	00:05:00	282	3.79	56.4
ES (F)	English	00:05:48	275	2.92	47.41
MS (F)	English	00:03:05	159	3.62	51.62
SJ (M)	English	00:05:47	277	4.71	47.92
TOTAL		00:23:44	1156	-	-

English M3D-TED					
Average (SD) per speaker English M3D-TED			231 (± 57)	3.93 (± 0.66)	48.76 (± 5.26)
DL (M)	French	00:05:36	324	3.79	57.86
FB (F)	French	00:09:13	359	5.33	38.94
JP (M)	French	00:05:04	230	4.04	44.75
KF (M)	French	00:07:22	197	7.21	26.77
MD (F)	French	00:10:17	414	4.29	40.23
TOTAL French M3D-TED		00:37:32	1524	-	-
Average (SD) per speaker French M3D-TED			304 (± 81)	4.93 (± 1.25)	41.71 (± 10.03)
OVERALL TOTAL M3D-TED		01:01:16	2680	-	-
Average (SD) per speaker across entire M3D-TED corpus			268 (± 79.03)	4.43 (± 1.12)	45.24 (± 8.75)

Table 2.7: Descriptive statistics of manual co-speech gesture annotation in the M3D-TED corpus.

Figure 2.5 shows the percentage of referential to non-referential gestures in the two corpora. The data showed that of the 1156 gestures in the English database, 750 were non-referential in nature (64.88%), while 406 (35.12%) contained some aspect of referentiality. Similarly, in the French database, of the 1524 gestures annotated, 1010 were non-referential in nature (66.28%), while 514 (33.73%) were referential.

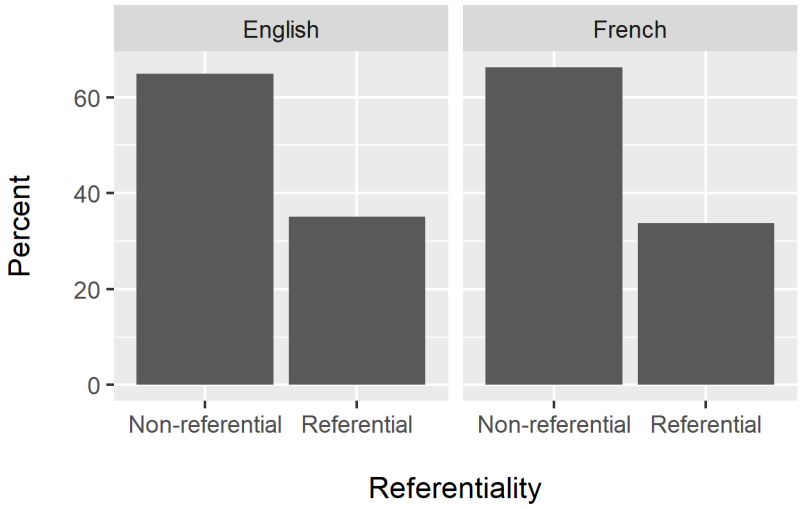

Figure 2.5: The percentage of referential and non-referential gestures in the English and French M3D-TED corpora

Figure 2.6 shows the different semantic categories that are the most frequently represented in the two M3D-TED corpora (expressed as the percent of occurrence relative to all of the semantic meanings labeled in the corpus). From the 406 referential gestures in English, a total of 476 different referential meanings were expressed. Similarly, of the 514 referential gestures in French, a total of 590

different referential meanings were expressed. The most common semantic meanings that were labeled were deixis and metaphoricity, followed by iconicity. The least numerous gestures were emblems.

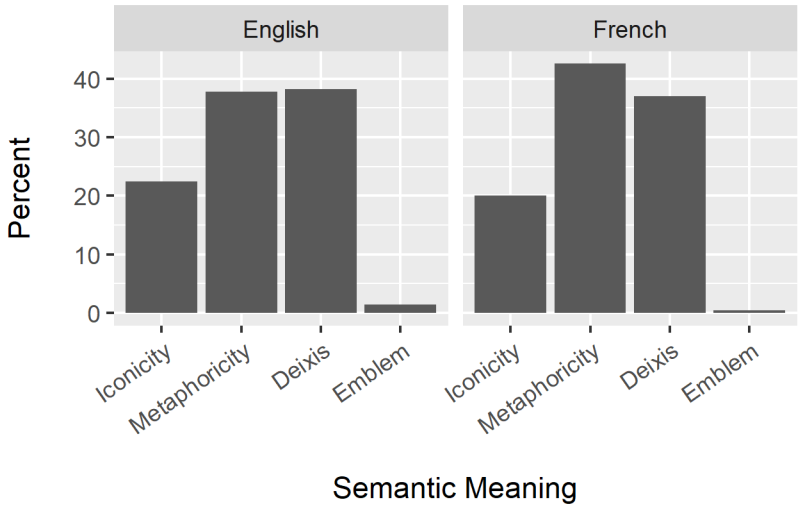

Figure 2.6: The percent of semantic (referentiality) categories represented by all gestures in each of the two M3D-TED corpora.

A total of 1113 gestural strokes from the English subset of the corpus were assessed audio-visually to see which pragmatic domains they could be expressing (see **subsection 2.3.3.5.2.**). The results of this preliminary coding are shown in **Figure 2.7**. They indicate that the gestures used by TED speakers express a number of pragmatic functions, particularly in terms of organizing discourse (72.6%) and expressing the speaker's stance (25.3%). Few gestures seem to be operational in nature or performing speech acts, and no gesture was reported to have an interactional function (which was to be expected, as the genre of speech is not dyadic). Moreover, there do not seem to be any contrasts between referential gestures and

non-referential gestures, suggesting that a "pragmatic/substantive" dichotomy may not be ideal. Of course, these preliminary results should be considered within the context of the genre of speech. For example, spontaneous conversation between two speakers will probably lead to more interactional (i.e., turn-taking) uses of gesture. However, these results illustrate how gestures contribute pragmatically to discourse, and suggest that using a comprehensive labeling system like M3D can enable further exploration of how their meanings are reinforced or mediated by other aspects of speech, such as prosody and morphosyntax.

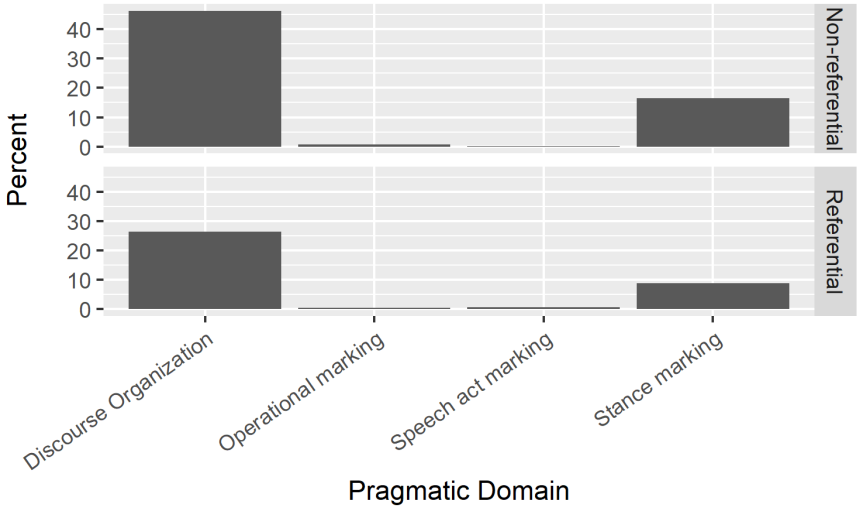

Figure 2.7: The distribution of pragmatic domains as a function of gesture referentiality.

2.3.6. Reliability of the key aspects of the M3D annotation

M3D's reliability was assessed by comparing independent annotations from the first author and a research assistant on approximately five minutes of speech in the English M3D-TED corpus (which represents ~20% of the entire English database). Four specific aspects of the annotations were chosen to be assessed for reliability as they represent the key aspects of gesture annotation and the most novel aspects introduced by M3D. Specifically, two aspects of the prosodic dimension were assessed, namely gesture phasing (i.e., the adequate segmentation and labeling of the different gesture phases, see **subsection 2.3.3.3.**), and apex placement. Additionally, two aspects of the meaning dimension were assessed, namely gesture referentiality and pragmatic domain annotation as non-mutually exclusive categories. Due to the nature of the coding system, reliability was assessed through different means depending on which aspect was to be assessed. The procedures and analyses will be detailed in the following subsections. In general, the present section will describe the English M3D-TED corpus elaboration procedure and reliability results.

2.3.6.1. Preparation of the English M3D-TED corpus for reliability

The annotation of the English M3D-TED corpus was carried out by a two-person team. The author of this thesis was in charge of training the second member of the team both in terms of the theoretical background (i.e., foundations in gesture studies) as well as specifically to annotate with M3D. Training and corpus

development took place over a 4-month period from January to April 2021 and consisted of biweekly meetings divided into three phases: an initial training phase, an annotation phase, and a reliability phase. The initial training phase consisted of approximately 12 sessions (over six weeks). These training sessions covered a variety of topics including the use of pertinent software, theoretical background on gesture, prosody, semantics and pragmatics, the M3D labeling system (including live joint coding sessions), and discussions on ambiguous cases, individual differences between speakers, and initial reliability.

The actual annotation phase continued over the following six week period, where bi-weekly meetings continued to take place to discuss ambiguous cases and evaluate and propose changes to the M3D system. During the annotation phase, only a select number of tiers were annotated, namely the gesture movement phasing tier, the apex tier, and gesture semantics (i.e., referentiality) tiers. Other aspects of annotation that are included in the M3D-TED corpus (namely ToBI prosodic annotations, annotations of information structure, and preliminary annotations of pragmatic meanings conveyed by gestures) have taken place outside of this main annotation effort, and will not be further commented here.

The final reliability stage lasted over two weeks where reliability between the two coders was evaluated for the English subset of the M3D-TED corpus. Following Kita et al. (1997), a general protocol involving two passes was followed, where coders independently

annotated approximately one minute's worth of gesture annotations per TED Talk. Initial reliability (i.e., the first pass) was then assessed based on the independent coding. Then, the two coders revised their annotations together so that cases of disagreement could be reviewed. A second assessment of reliability was then carried out over the same dataset (i.e., the second pass) to understand the extent to which the coders could not resolve their disagreements. The reliability results are reported for both passes in terms of gesture phasing, and at first pass for apex, semantic, and pragmatic annotations.

2.3.6.2. Gesture phasing annotation reliability

The first aspect we will discuss is movement phasing, which was assessed using ELAN's Calculate Inter-Annotator Reliability function to calculate Cohen's Kappa. This ELAN function is based on the algorithm created by Holle & Rein (2015), which works by "linking" annotations that overlap a minimum of 60% and then calculating the Cohen's Kappa value taking both segmentation and assigned value into account. **Table 2.8** shows the Cohen's Kappa values obtained across speakers for each potential movement phase (preparation, stroke, incomplete stroke, hold, and recovery).

The results of the reliability calculations showed substantial levels of agreement already at the first pass for all individual gesture phases as well as globally in terms of gesture phasing. As expected, the Kappa values increase after revising the annotations (see **section 2.3.6.1.**), showing very high levels of agreement in all aspects of

movement phasing. This suggests that the coders' independent annotations showed substantial agreement at the first pass, and furthermore, they were able to resolve most of their disagreements in terms of segmenting and identifying continuous streams of gestural movement.

	Cohen's Kappa Values	
	First Pass	**Second Pass**
Preparation Identification	0.7683	0.8914
Stroke Identification	0.7709	0.8846
Incomplete Stroke Identification	0.7994	1.0
Hold Identification	0.7817	0.8769
Recovery Identification	0.8196	0.9015

Table 2.8: The Cohen's Kappa values obtained across speakers for individual movement phases.

2.3.6.3. Apex annotation reliability

The second aspect we will describe is that of apex annotation placement. As the apex refers to a single point in time, and annotation is based on a frame-by-frame analyses, a qualitative assessment of reliability was used which evaluates the distance (in terms of 33ms video frames) coders placed apexes within linked

gesture strokes (as per Kita et al., 1997). As such, the aim of this reliability is to determine the reliability of such a frame-by-frame annotation of the apex, and less the omission/commission errors of apex annotation between the two coders. In other words, this reliability measure does not test whether both annotators identified an apex or not, but rather when both coders saw an apex, how closely in time did they code that apex.

Figure 2.8 shows the distribution of Coder 2's apex annotations relative to coder 1's annotations across time, binned in 33ms groups (i.e., one frame). Thus coder 1's apex annotation always occurs at 0ms, any annotation made by coder 2 that falls within a time range of -33 to 33ms would be considered as occurring within one frame (e.g., a distance less than that of 33ms). We see that of the 305 total apex pairs that were analyzed, the majority of the apexes coded by the second coder fall within one frame (33 ms) of the first coder's apex annotation (50.5%) and almost 73.8% of apexes were within two frames (66 ms) of each other. None of coder 2's annotations occurred more than three frames before coder 1's, and only 23 annotations occurred more than four frames after coder 1's. As such, the qualitative assessment of the reliability of coding the apex seems quite high, especially considering Loehr (2004) considered up to six frames of distance as being acceptable for coder agreement.

	-99 ms	-66 ms	-33 ms	0 ms	33 ms	66 ms	99 ms	123 ms
	3 frames	2 frames	one frame		2 frames	3 frames	4 frames	5+ frames
Raw Count	4	8	44	110	63	33	20	23
Percent	1.3%	2.6%	14.4%	36.1%	20.7%	10.8%	6.6%	7.5%
			50.5%					
		73.8%						

Figure 2.8: The distribution of Coder 2's apex annotations relative to coder 1's annotations across time, binned in 33ms groups (i.e., one frame)

2.3.6.4. Semantic annotation/referentiality reliability

The third aspect we will describe is that of gesture referentiality. For this aspect, Cohen's Kappa is not a suitable measure for reliability for two reasons: labelers could have multiple annotations for the same gesture (violating the assumption of mutual exclusivity), and the large number of labels for non-referential gestures could potentially result in the "Kappa Paradox" (a scenario where one label is observed significantly more than any other, affecting the calculation of chance agreement, e.g., Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Krippendorff, 2004).

To overcome this, Gwet's Agreement Coefficient 1 (AC1; Gwet, 2008) was calculated in R (R Core Team, 2021) with MASI (Measuring Agreement on Set-valued Items) as the distance metric (see, e.g., Artstein & Poesio, 2008; Passonneau, 2006). MASI distances quantifies the relative degree to which labelers agree for a set of non-mutually exclusive combinations of labels, while Gwet's AC1 uses the same formula as the Kappa yet calculates the chance

agreement taking this bias into account. Thus, it is resistant to the Kappa Paradox, yet can be interpreted in a similar fashion (see, e.g. Dettori & Norvell, 2020). In terms of the reliability for gesture referentiality, the global AC1 value across all five speakers resulted in very high rates of agreement (AC1 = .895, CI (.856, .933), $p < .001$).

2.3.6.5 Pragmatic annotation reliability

The fourth and final aspect to be discussed is the annotation of the pragmatic domain (for a total of four categories). For similar reasons described above, the calculation of reliability for the pragmatic domain followed the same analysis laid out for gesture referentiality, that is, Gwet's AC1 with MASI distance as the distance metric. The resulting global AC1 value again reveals a high rate of agreement between the labelers (AC1 = .78, CI (.726, .825), $p < .001$).

Based on various reliability metrics presented here, it seems that it is possible for coding teams to work together to achieve acceptable levels of agreement, particularly in terms of recognizing gesture strokes and phasing, precisely pinpointing the placement of the gesture apex, as well as understanding the semantic and pragmatic contribution of gestures to speech. While the reliability metrics presented here do not cover all aspects of M3D, the results from key novel aspects of M3D so far seem promising. Further comments and future steps will be further elaborated in the final discussion section.

2.4. Discussion and conclusions

The MultiModal Multi-Dimensional (M3D) Labeling system is a set of conventions for annotating multimodal speech corpora that incorporates not only the form properties of gestures (involving configuration and kinematic properties of multiple articulators), but also their prosodic, and meaning dimensions. It is based around the core idea that speech is multimodal in nature, and that these different modes of communication are not made up of mutually exclusive categories, but rather should be assessed in a holistic fashion. The main motivation to propose M3D was the current need within research in multimodal communication for a more standardized approach to multimodal data annotation that (a) incorporates recent advances in the gesture field regarding the multidimensional analysis of gestures; and (b) allows for annotations to be interpretable across investigation sites, and flexible enough to meet individual researchers' needs. As previously mentioned, current multimodal annotation systems vary widely in how they approach gesture taxonomies. And while some existing systems acknowledge the importance of coding multiple aspects of speech in multimodal recordings in order to understand interactions between different modes of communication, these systems do not systematically cover the three dimensions of gesture, namely gesture form, gesture prosodic features, and gesture meaning. Importantly, M3D disentangles gesture form from semantic and pragmatic meanings, as well as prosodic characteristics, proposing a tripartite dimensionality that is to be assessed for all gestures. Such an approach essentially bridges the

gap between two of the most prominent theoretical approaches in the field of gesture studies (i.e., Kendon, 2004; McNeill, 1992). Furthermore, in line with recent advances in the gesture studies field, M3D adds two important novelties: a) the assessment of semantic meaning in terms of gesture referentiality and the possibility of coding referential gestures in terms of potentially-overlapping dimensions rather than mutually-exclusive categories (in accordance with McNeill's (2006) proposal) and b) the possibility of annotating and thus exploring a range of non-mutually exclusive pragmatic meanings in a reliable manner. Through measures of inter-annotator agreement, we have shown that these key parts of the system can be reliably coded.

As previously mentioned, M3D also calls for the annotation of different modes of communication and aspects of speech to better understand interactions between them. Such an approach is crucial, as it allows researchers to elucidate complex relationships between, for example, gesture, speech prosody, and pragmatics. Importantly for the subsequent chapters in the current thesis, this approach will clarify the complex relationship between gesture and pitch accentuation by taking prosodic phrasal structure into account. Moreover, **Chapter 5** will shed light on the complex relationship between pitch accentuation, gesture, and the marking of information structure.

Summarizing, M3D represents an ongoing and evolving project which is aimed at helping researchers code multimodal data in a

standardized manner, through the inclusion of a detailed annotation guide, ELAN template, training materials, and a multimodal annotated corpus openly available online. Currently, M3D has been applied to the two M3D-TED corpora which are composed of 10 TED Talks (five in the English M3D-TED corpus, and five in French M3D-TED corpus) representing roughly in total over 60 minutes of multimodal data. Parts of the M3D have also been applied to the Audiovisual Corpus of Catalan Children's Narrative Discourse Development[11] (Vilà-Giménez et al., 2021). Specifically, the corpus has labeled aspects of the prosodic dimension, coding the rhythmic, phasing, and phrasing properties of gesture, as well as aspects of the meaning dimension, coding for referentiality and the gestural marking of information structure. Further development is planned for other genres and styles, such as spontaneous speech, settings with multiple interlocutors, and other forms of experimental multimodal data.

By offering standard annotation practices that are readily available in open access, M3D enables researchers across the domain of multimodal communication to have reliable and comparable results, hopefully fostering a more multidisciplinary and multidimensional approach for annotating the meaning-bearing elements of spoken language. Thus, we encourage the application of this labeling system to other multimodal corpora, and look forward to its continued development.

[11] https://osf.io/npz3w/

3

CHAPTER 3: PHRASAL PROSODIC STRUCTURE AS A KEY FACTOR IN THE TEMPORAL EXECUTION OF GESTURE IN FRENCH ACADEMIC DISCOURSES

3.1. Introduction

Research on the relationship between gesture and prosody has generally concluded that there is a tight temporal association between gestural prominence and prosodic prominence, with the two phenomena co-occurring at rates as high as 80% (e.g., Shattuck-Hufnagel & Ren 2018, among others). However, most gesture-speech alignment research has focused on languages where pitch accentuation generally has a prominence-lending function. Much less is known about the temporal alignment patterns in French, where pitch accentuation mainly serves a demarcative function, indicating the edges of prosodic phrases. To our knowledge, no study has explicitly assessed whether gesture production is modulated by phrasal position of the pitch accent (i.e., phrase-initial vs. phrase-final). Furthermore, while many studies have investigated the temporal association between gestures and prosodic prominence, less research has been devoted to the rhythmic production of subsequent gestures. The only two studies to our knowledge to have assessed rhythmic productions of subsequent gestures (i.e., McClave, 1994; Loehr, 2007) have found that gestural tempo seems independent of the tempo of subsequent pitch accents in speech. Taken together, these findings suggest that when multiple gestures are produced rhythmically, speech prominence may play a smaller role in temporal association of individual gestures with prominence and the tempo of subsequent gestural production. However, no previous study has explicitly compared referential gestures (those which convey semantic meaning) and non-referential gestures (those which do not convey any semantic

meaning, yet have been ascribed “rhythmic” characteristics). Furthermore, studies on the rhythmic production of subsequent gestures have only focused on pitch accentuation as the main rhythmic landmark in speech. All in all, less is known about how prosodic phrasing may influence the production of gesture.

To respond to these gaps, a corpus analysis was carried out on the French M3D-TED corpus, which contains over 37 minutes of Multimodal speech. The objectives of the current study are to assess the temporal relationship between gesture and pitch accentuation in French, the effects of referentiality on the rhythmic production of gesture, and the influence of prosodic phrasing on gesture tempo. Results showed that (a) gestures associate with pitch accentuation at rates similar to that of English, with a preference for phrase-initial positions when an initial accent is present (b) while non-referential gestures tend to be produced in a rhythmic fashion more often, gesture referentiality has no significant effect on the isochronicity, and (c) the duration of prosodic phrases significantly predicts the distance between subsequent gestures, where larger inter-onset-intervals in phrasing predict larger distances between subsequent gestural apexes. The results of the current study shed light on the impact of prosodic phrasing on gesture production, namely in terms of the temporal association patterns of pitch accentuation and gesture, as well as the complex relationship between rhythm in gesture and in speech in French.

3.1.1. The temporal association between prominence in speech and gesture

Speakers make use of multiple modes of communication to convey meaning in face-to-face interactions. Two such meaning-making strategies that are key in multimodal communication are speech prosody (i.e., intonation, rhythm, melody, etc.) and co-speech gestures (i.e., bodily movements that act as an utterance, Kendon, 2004). Furthermore, these two modes are closely related, in that prominence in gesture (i.e., the gesture stroke or apex) and prominence in speech (i.e., pitch accented syllables) tend to co-occur in close temporal synchrony (e.g., Loehr, 2004; Yasinnik et al., 2004; Jannedy & Mendoza-Denton, 2005; Leonard & Cummins, 2011; Esteve-Gibert & Prieto, 2013; Shattuck-Hufnagel & Ren, 2018; Pouw & Dixon, 2019b).

Recent research in the prosodic properties of co-speech gesture has made it increasingly clear that the traditional division between gestures which are referential in nature (showing degrees of iconicity, metaphoricity or deixis) cannot be prosodically distinguished from those which are non-referential in nature (i.e., "beat" gestures;" as per McNeill, 1992, 2006). While the former have been defined by their referential properties (pictorially representing semantic content or spatial relations via pointing), non-referential gestures have been traditionally defined as gestures which associate with prosodically prominent positions for discourse-pragmatic functions, which appear to be "beating musical time" (McNeill, 1992, p.15). However, recent studies have shown

that both referential and non-referential gestures associate with prominence in speech (instantiated through pitch accentuation) at similar rates. For example, Shattuck-Hufnagel & Ren (2018) investigated the overlap between gesture strokes and pitch accented syllables in academic lectures. They found that 82.85% of referential gesture strokes overlapped with a pitch accented syllable, which was not substantially different from the overlap rates of non-referential gesture (83.13%). Similarly, a kinematic analysis of gestural movements found no significant differences between gesture types for the distribution of various kinematic aspects of the gesture stroke (stroke onset, peak acceleration, peak velocity, and peak deceleration) and peak pitch (Pouw & Dixon, 2019b). Thus, gestures in general associate with prosodically prominent syllables.

Taken together, this body of research suggests that gestures associate with prosodically prominent positions in speech. However, much of the research on gesture-speech association has investigated English or other languages where pitch accentuation has a prominence-lending function (e.g., Italian, Catalan, Dutch, etc.) Importantly, preliminary evidence suggests that there may exist crosslinguistic differences in the association between gesture and speech prosody (see, e.g., Fung & Mok, 2018 for temporal association between pointing and speech in Cantonese, a tonal language) as well as in the rhythmic productions of gesture (e.g., Pouw & Dixon, 2019b). Thus, less is known about the relationship between gesture and speech when pitch accentuation is not considered to be prominence-lending. The French language offers

an interesting testing ground, as pitch accentuation is said to not be prominence-lending, but rather to have a demarcative function, delimiting prosodic phrase boundaries.

3.1.2. A brief overview of French prosodic structure

French is a fixed-stress language. Instead of being lexically distinctive (such as in English, where stress placement can distinguish a noun from a verb, such as in the common example of *REcord* vs. *reCORD*) stress placement is fixed, falling on the last non-schwa syllable of lexical words. Accentuation is then instantiated on the level of the smallest prosodic phrase. Though this phrase has received many names in the literature, the term Accentual Phrase (AP) will be used to describe this phrase in accordance with the Autosegmental-Metrical (AM) framework and French ToBI standards (Delais-Roussarie et al., 2015, see also Jun & Fougeron, 2000, 2002). Regardless of the theoretical approach to describe French prosody, it is widely held that this phrase contains at least one lexical word and all of the function words that it governs. The AM framework describes the AP as the level at which stress is cumulative. Specifically, pitch accents obligatorily mark the right edge of the AP, always occurring on the last full (non-schwa) syllable in the phrase. Thus, word final stress does not surface when lexical words are phrase medial. For example, the word *chaton* ("kitten" in French) will see a pitch accent on the final syllable when it is phrase final, but will not be pitch accented when phrase-medial (see **Example 3.1**). This obligatory phrase-final accent will be referred to as simply the Final Accent (FA). In

addition to the FA, a phrase-initial accent may optionally be placed on one of the phrase-initial syllables (henceforth "Initial Accent," or IA). As previously mentioned, accentuation is not lexically distinctive, so the two versions of the word *chaton* ("kitten") in **Example 3.1** (realized with FA, IA) are not lexically distinctive.

[3.1] (le cha**TON**)$_{AP}$ (ma**RRON**)$_{AP}$ *"The brown kitten"*
(le **CHA**ton ma**RRON**)$_{AP}$
Adapted from Delais-Roussarie & Di Cristo (2021)

The precise phonological status of the IA is still an area of some debate, where some researchers consider it a phrasal accent (i.e., that it merely associates with prosodic edges, e.g., Jun & Fougeron, 2000), a full pitch accent (i.e., a tonal event associating with a metrically strong syllable, e.g., Post, 2000), or a "hybrid" accent taking on characteristics of either depending on context (Grice, 2001; Portes et al., 2012). It can function to build up rhythmic patterns, breaking long stretches of speech that does not contain a pitch accent (e.g., Delais, 1994, as cited in Astésano et al., 2007; Janksowksi et al., 1999; Astésano, 2001), or be produced for emphatic or other pragmatic effects (e.g., Di Cristo, 1999, 2000). Regardless of its function, the IA is normally said to be marking the left edge of the AP (though its precise location may be variable, occurring on the first, second, or even third syllable of the AP) and can be realized on any class of words, be they lexical or functional (Astésano, 2001; Astésano et al., 2007; see Delais-Roussarie & Di Cristo 2021 for a review).

3.1.3. Studies on the temporal association between gesture and prosody in French

Only a handful of studies have assessed the relationship between gesture and prosodic structure in French. In a laboratory-based study, Roustan & Dohen (2010) asked participants to participate in a picture-naming task, where corrective focus was elicited. The experiment was carried out in separate blocks (conditions), where participants were instructed to point (pointing condition), produce a non-referential "beat" gesture (Beat condition), or press a button on the table (control condition). They found that when a word is prosodically focused and accompanied by a pointing gesture, a beat gesture, or a non-communicative movement (e.g., pushing a button), apexes "occur[ed] within or close to the focused element" (p. 4) and that the co-occurrence was closest with pointing gestures. Using the CID corpus of conversational speech, another study by Ferré (2010) investigated the temporal relationship between the *gesture phrase* of iconic gestures (i.e., the gesture stroke plus any related movement phases around such as preparations or holds, as per Kendon, 1980; henceforth, *G-Phrase*) and the Intonational Phrase as defined in Selkirk's (1978) Metrical theory (which according to Ferré, 2010, largely corresponds to the intermediate phrase in the AM approach). She reports that 70% of G-Phrases begin before the onset of the Intonational Phrase, and that 61% of G-Phrases end after the offset of the Intonational Phrase. In another study using the same corpus, Ferré (2014) investigated the association between gesture strokes of all types and marked structures in speech (i.e., syntactic fronting and prosodic emphasis).

She describes prosodic emphasis as the presence of an "unusually strong word onset" (p. 2) which would generally correspond to the realization of an IA (previously described). She showed that gestures reinforced prosodic emphasis more than marked (i.e., fronted) syntactic structures. Furthermore, she found that non-referential "beat" gestures associated with prosodic emphasis more than other gesture types. However, the study limits its investigation of gesture speech temporal association to "prosodic emphasis." No study to our knowledge has directly assessed the temporal association between pitch accentuation and gesture in French, controlling for phrasal position (IA vs. FA) and gesture type. Furthermore, fewer studies have specifically investigated the alignment properties of subsequent gestures that seem to be produced in a beat-like manner, appearing to mark speech rhythm.

3.1.4. Gestural rhythm

As previously mentioned, non-referential "beat" gestures have been characterized as closely associating with prominence and as "beating musical time" with speech (McNeill, 1992). Only a handful studies to our knowledge have empirically investigated such gesture production patterns and their rhythmic association with speech, using the gesture "apex" as the gestural landmark (described as the point of maximum excursion or "peak" of movement within the stroke, see Loehr, 2004, 2007). McClave (1994) looked at groups of subsequent beat gestures in conversational speech and found that not all apexes within the group coincided with pitch accented syllables. Instead, they tended

to follow their own rhythmic pattern, where at least one apex within the group was associated with a nuclear pitch accent. In other words, beat gesture apexes seem to occur at regular intervals spanning from the nuclear pitch accent, regardless of whether they associate with pitch accented syllables or not.

Furthermore, she observed that the apexes in these groups of gestures tend to be more isochronous (i.e., produced at regular intervals) when they are composed entirely of beat gestures. These findings lend support to her rhythm hypothesis - namely that beat gestures "are rhythmically patterned by themselves" and that "this pattern is not dependent on speech but, rather, meshes with speech at specific places to push the speech rhythm forward." (McClave, 1992, p. 46, see also Hardinson, 2019 for a qualitative study supporting these findings). Loehr (2007) discovered similar findings when comparing the rhythmic patterns of manual gesture, head nods, and eye blinks to pitch accentuation. He found that each phenomenon (hand, head, eye blinks, and speech prominence) showed their own rhythmic pace, yet the average tempo across articulators is approximately 300 ms intervals. He thus hypothesized that these different rhythms may be anchored in and come together around this common reference tempo. Finally, one preliminary study by Pouw et al. (2020) has investigated how crosslinguistic differences may result in rhythmic differences in gesture production. The researchers employed wavelet analysis to investigate the time scale oscillations of gesture production by Spanish-English bilinguals doing a narrative retelling in each

language and found a significant difference across languages. In their database, gestures produced while speaking English occurred at faster time scale oscillations than gestures produced while speaking Spanish. However, the preliminary study was limited in that the researchers only had the kinematic (gestural) data available, without any audio recordings to compare against speech prosody.

3.1.5. Motivation and research questions

The aforementioned results suggest that when groups of gestures are produced in a rhythmic fashion, they operate largely independent of speech rhythm instantiated by pitch accentuation in English. Moreover, preliminary evidence further suggests that there may be crosslinguistic differences in the rhythmic production of subsequent gesture. These approaches have largely focused on prominence but have neglected the influence of prosodic phrasing. French offers a unique testing ground in the assessment of the relationship between complex prosodic structure and rhythmic groups of subsequent gestures, as pitch accents regularly occur at the right edge of prosodic phrases. Thus, the regular marking of AP boundaries by pitch accents is crucial for the building up of speech rhythm, and indicates that the domain of the AP is key for speech rhythm in French (e.g., Pasdeloup, 1992; Mertens, 1992; Delais-Roussarie, 1995; Astésano, 2001). Thus, a second aim of the current study is to assess whether the production of subsequent gestures which appear to be "beating a musical rhythm" (henceforth Rhythmic Groups of Gestures, or RGGs) is constrained temporally

to associate with pitch accentuation, phrasing, or are largely independent of French prosodic structure.

To sum up, no study to our knowledge has specifically investigated the role of pitch accentuation as a prosodic anchor for gestural production in languages where pitch accentuation is not prominence-lending, and specifically if IAs (be they rhythmic or pragmatic in nature) act as special attractors for non-referential gestures (which have also been traditionally defined by their rhythmic and pragmatic functions). Furthermore, few studies have focused on the production of RGGs. Those that have have offered anecdotal evidence of differences by gesture type, and have only focused on the rhythmic relationship with pitch accentuation. None has taken prosodic phrasing into account, which is a key element in French prosodic structure. Thus, the current study aims to respond to the following questions:

1. Does pitch accentuation continue to act as a prosodic anchor for gesture, regardless of their demarcative function? Is this relationship modulated by accent type (IA vs. FA) or gesture type (referential vs. non-referential)?

2. Do non-referential gestures have a tendency to form RGGs more than referential gestures, and are non-referential RGGs more isochronous than referential RGGs?

3. Do RGGs tend to mark subsequent APs in French, and is this relationship modulated by pitch accentuation (i.e., the presence of absence of IA)? If not, is this relationship sensitive to the duration of prosodic phrases?

We hypothesize that pitch accentuation will act as a prosodic anchor for gesture production, regardless of their demarcative function, showing similar tendencies to what has been described for English. In terms of gestural landmarks, we hypothesize that strokes will be largely aligned with pitch accented syllables (around 80%, as per Shattuck-Hufnagel & Ren, 2018). By contrast, apexes will largely align with pitch accented syllables as well, but this relationship may be more variable than strokes (e.g., Pouw & Dixon, 2019b). In terms of phrasal position, it is believed that gestures will associate more with the IA in French, as these accents may serve more pragmatic or emphatic functions than FA, which mainly function to delimit the right edge of the AP. Importantly, no differences between referential and non-referential gestures are expected to surface. Given the tendency for gestures to align with pitch accents, we hypothesize that RGGs will also closely mirror pitch accentuation, where there will generally be one gesture per AP, which may double when APs contain two pitch accents. Finally, we predict that both referential and non-referential gestures will be produced as RGGs, with no differences in isochrony as recent work has questioned the idea that certain gesture types are more closely related to rhythm and prosody (e.g., Shattuck-Hufnagel & Prieto, 2019).

3.2. Methods

3.2.1. Materials: The French M3D-TED Corpus

The French M3D-TED corpus was used in the current analysis. The audiovisual corpus contains over 37 minutes of multimodal speech from five different native adult Metropolitan French speakers giving a TED Talk (mean duration per speaker: 07m 30s). The corpus contains a total of 1524 gesture strokes, 1770 apexes, and 3912 pitch accented syllables. After removing stretches of silence or disfluent speech, a total of 1504 strokes and 1698 apexes remained in the database for analysis.

TED talks can be regarded as a form of academic speech. It has been described as a "hybrid genre" (Caliendo, 2012, p. 101, as cited in Matiello, 2019) - similar in format to a conference talk, yet the members of the audience are often not specialists. Consequently, a rather informal register is often adopted by TED speakers, making it more similar to spontaneous conversation (Mattiello, 2017; see Mattiello, 2019 for an overview). TED Talks are an ideal genre for the study of gesture, as TED speakers are generally quite expressive and a good number of gestures typically appear in TED Talks (see, e.g., Harrison, 2021). Specifically in the French TED Talk corpus, the mean rate of words per manual gesture, considering both referential and non-referential, is 4.93 (i.e., a gesture is produced approximately every five words on average). Though these talks are oftentimes rehearsed and/or trained, the official TED guide to public speaking (Anderson, 2016) does not give specific details on

how speakers should employ specific prosodic or gestural features in their speech. Rather, the guide proposes that speakers should speak naturally and conversationally. Specific advice regarding the use of prosody include using varied speech rhythm and intonational patterns that are coherent with the meaning the speakers wish to convey. Similarly, the guide proposes that speakers move their bodies (i.e., gesture) intentionally and make use of their hands and arms to amplify their message in speech. Importantly, the guide highlights that this should come naturally and that there are no "rules" the speakers should follow. For these reasons, we believe TED Talks can be classified as natural, academic style discourse.

3.2.2 Data annotation

The French M3D-TED corpus was independently annotated for prosody and gesture. The entire corpus is available online[12] in the format of ELAN files (Wittenburg et al., 2006), as well as the M3D labeling manual which explicitly describes the annotation procedure and each tier that is available in the corpus (Rohrer et al., 2021). The following subsections will describe the annotation tiers that are related to the current study.

3.2.3. Gesture annotation

Gestural annotation was carried out by the author of this thesis within the context of developing the MultiModal MultiDimensional labeling system, following the annotation guidelines which are fully described in the labeling manual (Rohrer et al., 2021; see also

[12] https://osf.io/ankdx/

Chapter 2). It makes use of the gesture phrasing and phasing tiers, the gesture referentiality tierset, and the assessment of RGGs.

3.2.3.1. Annotation of gesture phrasing and phasing

Specifically, only manual co-speech gestures were coded (that is, meaningful manual movements that act as an utterance, or part of an utterance, as per Kendon, 2004). All gesture annotation was carried out using frame-by-frame analysis in ELAN. Furthermore, as it is explained below, initial passes were carried out without access to the audio, so as to avoid influence from the speech stream (i.e., for the annotation of gesture units, phases, and apexes).

First, manual gesture units (G-Units) were identified and annotated, which corresponds to moments from when the hands leave a position of rest or relaxation to their subsequent return to rest or relaxation. Each G-Unit was then divided into the various gesture phases (preparation, stroke, hold, recovery). Stroke identification was largely based on the kinematic properties of the movement (salient movements based on speed, changes in handshape, etc.). The apex was annotated on a separate tier. The apex is described as any sudden stops, changes in direction or moments of zero velocity, which can be seen as the peak effort in the stroke (see, e.g., Loehr, 2004; Yasinnik et al., 2004). The apex is identified in frame-by-frame analysis as corresponding to the frame in which the hand(s) go from blurry to suddenly being clear (see **Figure 3.1**), or the frame immediately preceding one in which the direction of movement changes.

Figure 3.1: Still images of a (non-referential) gesture executed in the French M3D-TED corpus, by speaker JP (TEDx Talks, 2018) at 02:58. **Upper panel:** the various gesture phases involved in the execution of the gesture. **Lower panel:** frame-by-frame images of the stroke, where the final frame indicates the apex.

3.2.3.2. Gesture referentiality

After an initial pass to label gesture phrasing, phasing, and apex, a second pass was carried out with audio in order to assess gesture referentiality for each stroke. Following the guidelines set by M3D, gestures can be divided into referential and non-referential gestures. The former have a clear referent in speech through representation (degrees of iconicity or metaphoricity) or by showing spatial

relationships (deixis), while the latter do not have a clear referent in speech.

3.2.3.3. Annotation of the Rhythmic Groups of Gesture (RGGs)

The procedures described in McClave (1994) and Loehr (2007) were adapted in order to identify RGGs. RGGs specifically refer to the production of subsequent gestures that are perceived as "beating out musical time," that is, they are produced in a rhythmic fashion. To identify RGGs, the video was played at full speed (without audio). In order for a series of subsequent gestures to be considered as belong to a beat-like, "rhythmic group," it had to satisfy the following conditions:

1. RGGs had to appear to be "beating musical time" (i.e., have a certain rhythmic quality).

2. Any potential RGG must contain at least three apexes (so as to be able to measure at least two inter-onset-intervals and assess isochrony).

3. RGGs could not cross gesture unit boundaries, contain major changes in hand configuration or trajectory.

Following these criteria, RGGs could consist of all of the gestures within a G-Unit, or only a select number of gestures. Additionally, hand configuration and trajectory should be similar across the gestures within an RGG. Importantly however, the referentiality

properties of the gesture was not considered when assessing the rhythmic quality. Therefore, the RGGs could be made up of entirely non-referential gestures, referential gestures, or a mix between them (i.e., alternating between non-referential and referential gestures). **Figure 3.2** shows an example of the gestural annotation for an RGG in ELAN. The RGG was composed of three non-referential strokes co-occurring with the utterance "En fait, ces hommes et ces femmes" (*in fact, these men and women*). The upper panel shows the still-frame images at the onset and apex of each of the three gesture strokes forming the RGG. The lower panel shows the annotations in ELAN, where the first tier represents the gesture phasing (preparation/stroke), the second tier represents the apex (endpoint of the interval annotation), and the third shows the orthographic annotation (words). The fourth and fifth tiers represent the prosodic information, namely the boundaries of the pitch accented syllables as well as the pitch accent type (Syll_tones) as well as the boundaries of each AP (see **subsection 3.2.4.**). Finally, the arrows between the two patterns indicate correspondence between the still images and their location in the ELAN annotations. **Figure 3.3** shows another example of an RGG, but with a referential gesture exhibiting deixis.

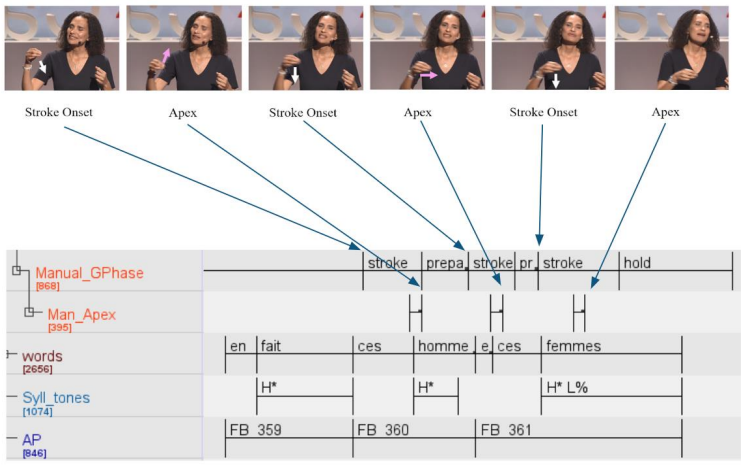

Figure 3.2: Example of RGG production with the utterance "En fait, ces hommes et ces femmes" ("In fact, these men and these women…") from the French M3D-TED corpus, by speaker FB (TEDx Talks, 2015) at 06:54. **Upper panel**: Still-frames extracted at stroke onset and apex for each of the three gestures (arrows indicating direction of upcoming movement). **Lower panel**: ELAN annotations of the RGG, including gesture phrasing, apex annotation, words, pitch accented syllables, and prosodic phrases.

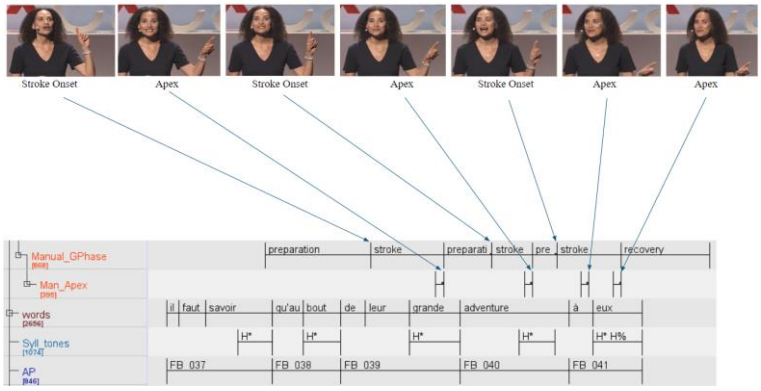

Figure 3.3: Example of RGG production with the utterance "il faut savoir qu'au bout de leur grande aventure à eux" (*You should know that by the end of their own big adventure*) from the French M3D-TED corpus, by speaker FB (TEDx Talks, 2015) at 01:22.

3.2.4. Prosodic annotation

Prosodic annotations were carried out by the author of this thesis. An orthographic transcription of speech was initially carried out in Praat (Boersma & Weenink, 2022). The transcription was then automatically aligned and segmented into words, syllables, and phones with the Montreal Forced Aligner (McAuliffe et al., 2017). Prosodic labeling was then carried out following the French ToBI (Tones and Breaks Indices) system (Delais-Roussarie et al., 2015). Two main domains were labeled, namely phrasing and pitch accentuation. Regarding prosodic phrasing, a breaks tier was used to assess phrasing across five levels: a 0-break indicates a grammatical word boundary, a 1-break indicated a lexical word boundary, a 2-break indicates an AP boundary, a 3-break indicated an intermediate phrase (ip) boundary, and a 4-break indicates an Intonational Phrase (IP) boundary. Of particular importance for the current study is the annotation of the AP-boundaries.

Regarding pitch accentuation, a tones tier was used to assign the tonal target to pitch accented syllables, as well as phrasal accents (at ip boundaries) and boundary tones (at IP boundaries). For pitch accentuation, polysyllabic words with a pitch movement on the right edge were generally labeled as an FA, while polysyllabic lexical words with a pitch movement on the left edge were labeled as an IA. Monosyllabic words which contain pitch rises are often ambiguous as to whether these are initial or final accents. In such cases, a conservative approach was taken, generally marking them as FA (particularly when the word was lexical). **Figure 3.4** shows

an example of the prosodic annotations carried out in Praat which correspond to the example illustrated in **Figure 3.2**, where the first tier indicated the word boundaries, the second tiers indicates syllable boundaries, the third tier indicates pitch accentuation, the fourth tiers reflects breaks, and the final two tiers reflect the two levels of phrasing (APs and IPs). Once the prosodic annotations were completed in Praat, the annotations were imported into ELAN. The gestural and prosodic annotation data was then exported together in a time-aligned database for further processing in R (R core team, 2021).

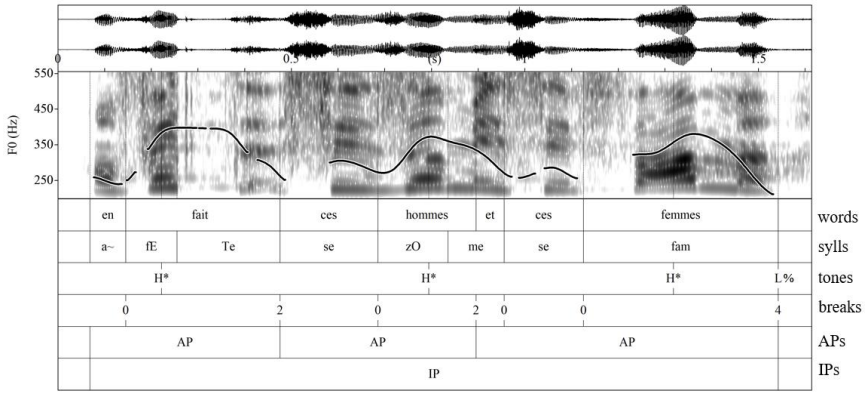

Figure 3.4: Example prosodic annotation for the utterance "En fait, ces hommes et ces femmes…" (*In fact, these men and these women…*) from the French M3D-TED corpus, by speaker FB (TEDx Talks, 2015) at 06:54.

3.2.5. Gesture-speech alignment criteria

In order to assess the temporal association of gestures with speech, the temporal overlap between prosodic and gestural landmarks was assessed. While the prosodic landmark of interest was the temporal

span of pitch accented syllables, two key gestural landmarks were assessed: the stroke phase, and the apex. First, each stroke was assessed for whether it overlapped with a pitch accented syllable or not. In other words, if any part of the stroke annotation temporally occurred within any part of the annotation of a pitch accented syllable, then the stroke was considered to have aligned with a pitch accented syllable (e.g., Shattuck-Hufnagel & Ren, 2018). Apexes were also assessed for whether they aligned with pitch accented syllables (i.e., if the point in time that refers to the apex fell within the boundaries of a pitch accented syllable, it was considered as aligned).

3.2.6. Statistical analyses

In order to assess the relationship between pitch accentuation and gesture production, descriptive statistics were carried out to assess temporal alignment between gesture strokes and apexes and pitch accented syllables. To assess whether the relationship between strokes and pitch accented syllables was affected by pitch accent type or gesture type, a Generalized Linear Mixed-effects Model (GLMM) was run using the *lme4* package (Bates et al., 2015). The model was run with the number of strokes as the dependent variable, with a fixed factor of Accent Type (2 levels: IA and FA), a fixed factor of gesture type (2 levels: Referential and Non-referential), as well as their two-way interaction. In order to assess the best random effects structure that fits the data, the *buildmer function* (Voeten, 2022) was applied, which compares all potential combinations of random effects and returns the best fitting model.

The function suggested a random effects structure of random slopes and intercepts by Speaker[13]. Omnibus test results were then carried out to assess significant main effects, which were assessed with a series of Bonferroni pairwise tests carried out with the *emmeans* package (Lenth, 2022).

To determine whether RGGs tend to be more referential or non-referential in nature, a Generalized Linear Mixed-effect Model (GLMM) with a poisson regression was run, with the Number of RGGs as a dependent variable, and a Fixed Factor of RGG referentiality (2 levels: Referential and Non-referential). The model included random slopes and intercepts by speaker, and was offset by the total number of gestures produced by referentiality[14]. The assessment of isochronicity between RGGs as a function of their referentiality was based on the calculation of the normalized Pairwise Variability Index (nPVI; Grabe & Low, 2002). This measure was chosen as it is a relatively standard rhythm metric in linguistics and offers a simple numerical value between 0 to 200, where values closer to 0 indicate less variation between Inter-Onset-Intervals. In order to assess difference in isochronicity between RGG referentiality types, a Linear Mixed-effects Model (LMM) was run with nPVI as the dependent variable, RGG referentiality as

[13] glmer(data = df, N_strokes ~ AccentType*GestureReferentiality + (1 | Speaker), family="poisson")
[14] glmer(data = df, N_RGGs ~ RGGReferentiality + (1 | Speaker), offset = log(TotalGestures), family="poisson")

a fixed factor (3 levels: Referential, Non-referential, and Mixed), and random slopes and intercepts by speaker[15].

To assess the correspondence between the RGGs and prosodic phrases in terms of pitch accentuation, a Generalized Linear Mixed-effect Model (GLMM) with a poisson regression was run, with the Number of occurrences as a dependent variable, and a Fixed Factor of AP configuration (2 levels: 1 pitch accent or 2 pitch accents), a Fixed Factor of apex number (4 levels: 0, 1, 2, or 3 apexes), and their two-way interaction. The model included random intercepts for AP configuration by speaker, and was offset by the total number of APs produced by configuration[16]. Finally, to assess the relationship between Apex IOI and AP IOI, a simple linear regression was run with average apex IOI as a function of the average AP IOI within the RGG[17].

3.3. Results

3.3.1 Temporal alignment between manual gesture strokes and apexes with pitch accented syllables

In response to the first research question, one of the main goals was to assess whether pitch accentuation continues to act as a prosodic anchoring point for gesture production (focusing on two landmarks, namely the stroke and the apex) regardless of its demarcative

[15] lmer(data = df, nPVI ~ RGG_Referentiality + (1 | Speaker))
[16] glmer(data = df, N ~ AP_configuration*N_apexes + (1 + AP_configuration | File), offset = log(Total_AP), family = "poisson")
[17] lm(AvgApexIOI ~AvgAPIOI, data=df)

function. **Table 3.1** below shows the by-speaker comparisons for both levels of temporal alignment. Alignment patterns for strokes were very similar to what has been previously reported in the literature for English. Specifically, the average rate of alignment between strokes and pitch accented syllables was shown to be 90.32% (SD: 4.69%). However, alignment rates for apexes were much lower than expected, occurring within the pitch accented syllable at an average rate of 50.8% (SD: 4.81%).

Speaker	Stroke Alignment (%)	Apex Alignment (%)
DL	92.21%	56.17%
FB	84.29%	47.48%
JP	94.3%	52.03%
KF	85.2%	43.29%
MD	95.59%	55.02%
OVERALL	90.69%	51.53%

Table 3.1: The alignment rates between gestures and pitch accented syllables in the French M3D-TED corpus, separated by speaker (Column 1), and between gesture strokes and apexes (Columns 2 & 3).

The second objective of this research was to assess whether accent type (i.e., AP-initial vs. AP-final accents) act as stronger attractors for gesture-speech association. For this analysis, 748 gesture strokes were removed from the analysis as they either overlapped multiple pitch accents (N = 596) or did not overlap any pitch accented syllable (N = 141) or because they occurred during disfluent speech (N = 12). An initial inspection of the data (N= 776) suggests that gesture associates with FA more than with IA. However, this is likely due to the abundance of FA in the database. When looking exclusively at gestures which align with pitch accents in APs which contain both potential anchoring points (i.e., which contain both an IA and an FA; N=257), the GLMM revealed a significant main effect of Pitch Accent type, ($\chi^2(1) = 31.9$ $p < .001$), indicating that gesture occurred significantly more with IA than with FA ($z = 4.98$, $p < .001$). A significant main effect of gesture type was found as well ($\chi^2(1) = 36.97$ $p < .001$), indicating that there were significantly more non-referential gestures than referential ones ($z = 5.446$, $p < .001$). However, no significant interaction between the two was found ($\chi^2(1) = 0.001$ $p = .978$) (see **Figure 3.5**). Taken together, these results indicate that when two potential anchoring points are available within the AP, gestures associate with the AP-initial accent.

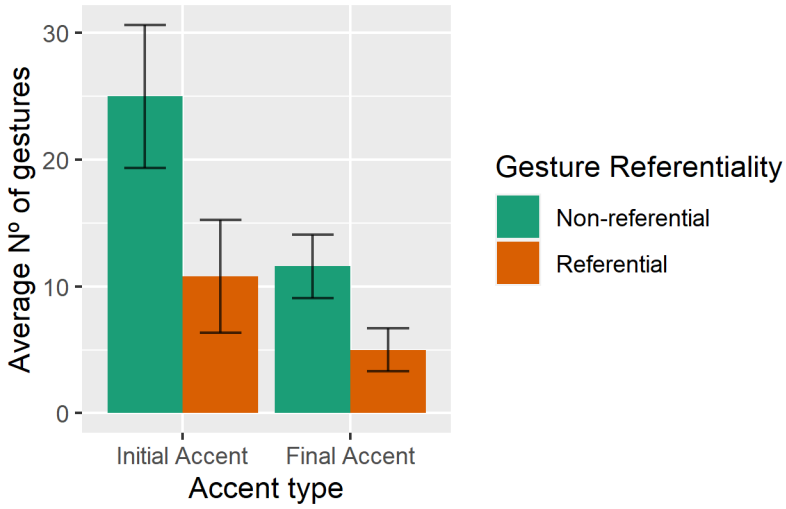

Figure 3.5: The average number of gestures per speaker as a function of their alignment with IA or FA by gesture type (error bars represent standard error).

The second and third research questions pertain to the rhythmic production of subsequent gestures (i.e., RGGs). In order to assess this aspect of gesture production, the gesture apex was chosen as the gestural landmark for several reasons. While it was shown in **subsection 3.3.1** that the apex does not reliably fall within the bounds of a pitch accented syllable, a closer inspection of non-aligned apexes revealed that they most often occur on the syllable immediately preceding or following a pitch accented syllable. Thus, they are still within close temporal proximity to pitch accented syllables. Additionally, the apex refers to a single point in time, as opposed to an interval. As the basis for rhythmic analysis in the current study is the Inter-Onset-Interval (IOI), the interval between subsequent apexes is a more precise measure of "maximum effort" (as per Loehr, 2004; 2007) and is less affected by potential differences in, e.g., stroke duration, where onset/offset of strokes

may be more variable. Specifically for the third research question, the bounds of the AP were of interest. This prosodic landmark was chosen as the right edge regularly occurs with a pitch accented syllable (see **subsection 3.1.2.**) and previous studies have shown it to be important in the perception of speech rhythm (e.g., Astésano, 2001; 2017).

3.3.2 The rhythmic productions of referential and non-referential gestures

The second research question addresses whether non-referential gestures tend to be produced in a rhythmic fashion (i.e., are more likely to form RGGs) more than referential ones, and whether non-referential RGGs tend to be more isochronous than Referential ones. The database contained a total of 183 RGGs (containing a total of 629 gestures), of which 106 were composed entirely of non-referential gestures, 19 of referential gestures, and 58 contained a mix of both gesture types. Removing the 58 RGGs which contained both referential and non-referential gestures, the GLMM revealed a significant main effect of Gesture Referentiality ($\chi^2(1) = 11.249$ $p < .001$), indicating that RGGs tended to be more non-referential in nature than referential ($z = 3.354$, $p > .001$). Thus, it seems that subsequent non-referential gestures are particularly perceived to be produced in a rhythmic manner. However, when assessing the isochronicity of the RGGs by referential types, the model revealed no significant differences between RGG referentiality ($\chi^2(2) = 1.096$ $p = .578$), indicating that non-referential RGGs are not more isochronous than referential ones (see **Figure 3.6**).

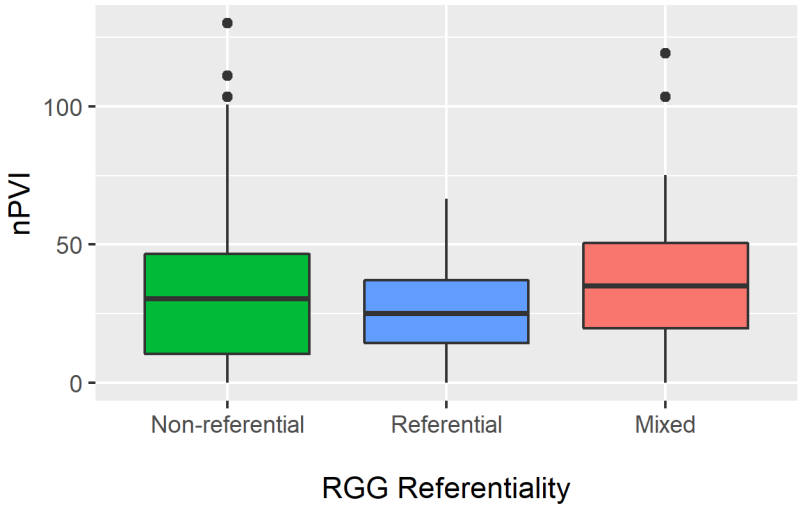

Figure 3.6: The nPVI of RGGs, as a function of their referentiality.

3.3.3. The relationship between RGGs and APs

The final research question addresses the relationship between RGG apexes and APs. Specifically, we were interested to see how many RGG apexes generally fall within each concurrent AP, and if apexes marked subsequent APs so that each AP co-occurring with an RGG regularly received one apex. **Figure 3.7** (left panel) shows the frequency of the number of RGG apexes within each individual AP. Of the 708 APs that were co-produced with RGGs, 532 (75.14%) were marked by a single RGG apex, while 79 (11.16%) did not contain an apex, and 89 (12.57%) contained two apexes. Only eight cases had more than two apexes. Thus overall, APs within RGGs tend to mostly receive a single apex. However, if we consider the marking of subsequent APs (that is, if the tendency is for each AP co-produced with a single RGG to receive a single apex), we see that of the 183 RGGs, only 73 (39.89%) follow a one-to-one pattern

(i.e., each AP receiving one RGG apex). Rather, 110 (60.12%) RGGs follow a non-one-to-one pattern, indicating that it is common for APs to be skipped or doubly marked by rhythmically produced subsequent gestures (**Figure 3.7**, right panel).

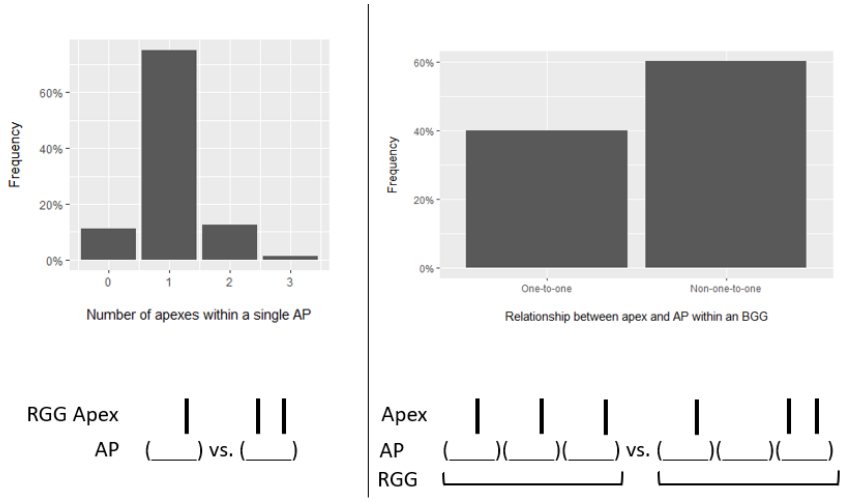

Figure 3.7: The relationship between RGG apexes and prosodic phrasing. **Left Panel:** The correspondence between RGG apexes and individual APs, showing the frequency of APs containing 0, 1, 2, or 3 apexes. **Right panel:** The correspondence between apex and APs concurring with RGGs, showing the frequency with which subsequent APs each contain a single apex (one-to-one) and frequency with which subsequent APs vary in terms of the number of apexes they contain (not-one-to-one).

These descriptive results suggest that APs that are co-produced with RGGs are not being marked in a regular manner, with some APs being skipped, while others within the RGG receive multiple apexes. However, the APs themselves are variable in terms of how many pitch accents they may contain, with some only having the

FA, while others may have an IA. Thus, it is necessary to assess whether this variability in rhythmic gesture production could be explained by pitch accent configuration within the AP (i.e., APs that contained two accents would lead to that AP also containing two RGG apexes). **Figure 3.8** shows the distribution of the number of apexes contained in the AP as a function of the number of accents in the AP. The results of the GLMM revealed a significant main effect of Apex number ($\chi^2(3) = 408.719$, $p < .001$) and a significant interaction between the Apex Number and AP pitch accent configuration ($\chi^2(3) = 45.765$ $p < .001$). The post-hoc pairwise comparisons of the significant interaction revealed that when APs did not contain a gesture apex (i.e., they were "skipped" in the rhythmic gestural marking), they were significantly more likely to contain only one pitch accent ($z = 5.88$, $p < .001$). Furthermore, APs with two accents were not significantly more likely to receive multiple apexes than APs with one accent, and APs were significantly more likely to receive a single gestural apex regardless of the number of pitch accents they contain. These results suggest that this variability in the co-production of speech and rhythmic gesture is not a direct result of pitch accent configuration.

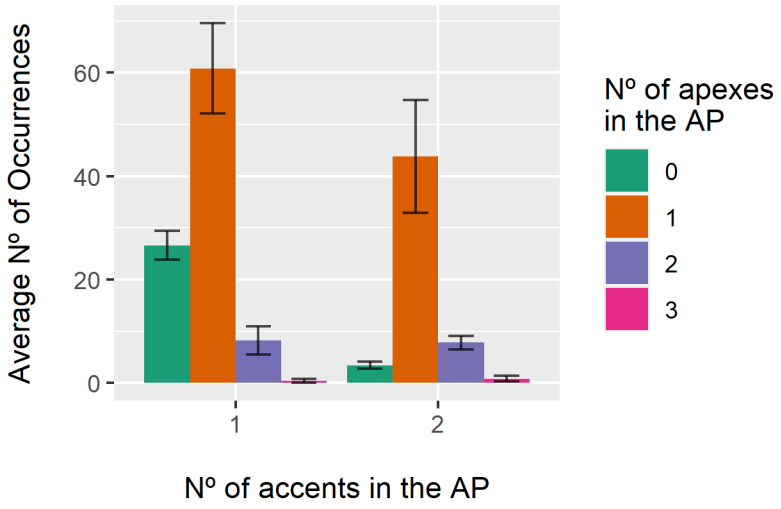

Figure 3.8: The distribution of the average number of apexes contained in the AP by speaker as a function of the number of pitch accents in the AP (error bars represent standard error).

The final analysis assessed whether or not the tempo of RGG is sensitive to prosody in terms of phrasing (specifically, the duration of APs). The linear regression showed a significant relationship between the average AP IOI and average Apex IOI (R^2 = 0.333, F(1, 162) = 82.42, p < .001). Thus, in general, as APs become larger, so does the interval between successive gesture apexes (see **Figure 3.9**), and this relationship explains approximately 33% of the variability in the data. Taken together, these results suggest a complex rhythmic relationship between speech and gesture.

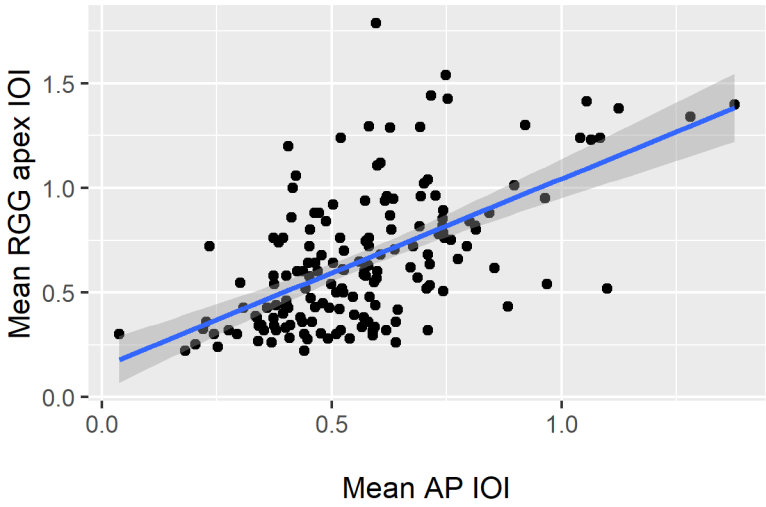

Figure 3.9: Scatterplot showing mean RGG apex IOI as a function of the mean AP IOI.

3.4. Discussion and conclusions

The aim of the current study was to assess gesture-speech temporal association in French. Specifically, one objective was to investigate whether pitch accentuation continues to serve as a prosodic anchor for gesture production in French, and whether there was any preference for gestures to associate with phrase-initial or phrase-final accents. The second objective was to investigate the rhythmic production of gestures, specifically asking whether non-referential gestures are inherently more rhythmic in terms of frequency and isochronicity. The final objective was to assess the rhythmic relationship between co-speech gesture and speech prosody at the level of the AP, assessing whether there is a regular correspondence between APs and RGGs, and if this relationship is sensitive to accent configuration or AP duration.

In terms of the first objective, we found that strokes align with pitch accented syllables to similar degrees as has been previously reported for English (e.g., Shattuck-Hufnagel & Ren, 2018). To our knowledge, this study is the first to investigate the alignment patterns in French in such a way to be comparable to previous studies in English. However, the alignment of apexes was much lower compared to what has been reported in previous literature. As previously described in most studies have focused on languages where pitch accentuation has a prominence-lending function such as English (Loehr, 2004; 2012), Catalan (Esteve-Gibert & Prieto, 2013) or Italian (Esposito et al., 2007). At this stage, it remains unclear if this may be a typological effect of language, as pitch accentuation in French is not prominence-lending. Further studies will need to assess the reliability of the apex as a landmark that closely associates with pitch accented syllables (see **Chapter 4** of this thesis).

Importantly, the current study further sheds light on the gesture-speech alignment patterns in French, finding that when both IA and FA are present within an AP, the gesture will align with the IA significantly more often than with the FA. If the presence of an IA simply offered a second equally potential anchoring point for gestures to align, then we would not expect a significant difference in alignment. In other words, IA and FA are different in their relative force for attracting gesture. Our results put into question a strict view of the one-to-one relationship between gesture

prominence and prosodic prominence, as in our data there is a clear preference for IAs over FAs when they are available, suggesting that not all prominences equally attract gesture.

One of the explanations as to why the IA is a stronger gesture attractor may be related to how it functions in speech with respect to the FA. The FA regularly marks the right edge of the AP (but which may be upgraded to nuclear status particularly in cases of narrow focus). Similarly, in addition to marking the left edge, IAs can sometimes be considered "emphatic" (separate from traditional IAs, as emphatic IAs are often realized with a longer syllabic duration, Astésano et al., 2007; see also Astésano, 2017). As these accents are more optional, it may well be that the speakers have chosen to produce this accent to some emphatic effect (particularly with TED Talks being academic and "inspiring" in nature), which may additionally trigger the production of co-speech gesture (in line with the pragmatic synchrony rule per McNeill, 1992, which holds that pragmatics in speech and gesture are coherent).

The fact that "emphatic" IAs are phonetically realized with a longer syllable duration (as mentioned above) suggests that these may also be more prominent. As such, the degree of relative prominence may also play into the role of IA as a gesture attractor. However, a study by Hualde et al. (2016) investigated the prominence ratings of IA in French conversational speech (without taking into account whether they were emphatic or rhythmic in function) and found that IAs tend to be perceived as less prominent than FAs. Thus, future

studies should disentangle the relationship between relative prominence and phrasal position of pitch accents to assess whether gestures associate with stronger prominences, or if structural factors (i.e., a "left-edge effect") are at play, where gestures are attracted to early positions in the phrase (see **Chapter 4** of this thesis).

The finding that gesture associated with IA also has implications in the field of prosody, as it may help disambiguate the status of the IA. Indeed, it has been described variously as a boundary tone, a pitch accent, and a "hybrid" tone that may function as both depending on context (see **subsection 3.1.2**). However, a number of studies have found that gestures associate more with pitch accents than with boundary tones (e.g., Esposito et al, 2007; Loehr, 2004; 2012; Turk, 2020). Thus, the results from the current study seem to suggest that the IA is not a boundary tone. While these results seem to suggest that it is indeed a pitch accent, it does not discard the possibility of it behaving as a boundary tone at times (i.e., the "hybrid" view). One avenue for future research may be to assess the pragmatic function of IA (i.e., whether the IA is rhythmic or pragmatic) and assess whether gesture occurs in all conditions or only certain. Doing so may advance our knowledge about the IA in French.

The second objective of this study aimed to investigate the rhythmic patterns of referential and non-referential gestures. Specifically, we asked if certain gesture types tend to be produced more frequently in a rhythmic fashion, and whether there were any differences in

isochronicity between them. First, we found that rhythmic, beat-like groups of gestures tended to consist entirely of non-referential gestures more often than referential ones. Importantly, the statistical analyses controlled for the overall production of gestures by referentiality, suggesting that this is not merely an artifact that more non-referential gestures were produced as a whole. These results thus lend support to the idea that non-referential gestures may group together to appear to be "beating musical time" of speech, as described by McNeill (1992, p. 15). It is quite likely that the perception of rhythmic productions of non-referential gestures is due to their lack of semantic content. Referential gestures (particularly those which pictorially represent semantic content in speech) make use of specific hand configurations and trajectories which must adequately portray semantic content. Greater complexity in form may have an impact on producing subsequent gestures in a rhythmic fashion. In contrast, non-referential gestures are not constrained by the need to represent semantic content. Without such constraints on form, they may lend themselves to being produced subsequently in a rhythmic fashion. The results from a study by Shattuck-Hufnagel & Ren (2018) also highlight the role of gestural form and its integration with prosodic phrasing. Though their analysis of gesture groupings were not temporal in nature, they grouped subsequent gestures by form (that is, subsequent gestures that showed very similar configuration and kinematic features were dubbed as "perceived gesture groups", or PGGs) and PGGs were found to align with higher levels of prosodic phrasing (above the level of the IP). Thus, future research should

assess how form features may be involved in complex relationships between gesture and prosody. In any case, it is important to note for the current study that not all non-referential gestures in this corpus were perceived as being a part of an RGG, thus it would be inaccurate to assign this prosodic characteristic to all non-referential gestures.

Importantly, when comparing the isochronicity of RGGs by type, the data showed no evidence that non-referential RGGs are more isochronous than referential ones. While this latter finding seems contradictory to those previously reported in McClave (1994), the author doesn't offer any quantitative analysis of isochrony or in-depth discussion to support her claim that non-referential RGGs tend to be more isochronous. Further, this finding is not entirely surprising as more and more research suggests that prosody and gesture are closely integrated regardless of referentiality (e.g., Shattuck-Hufnagel & Ren, 2018; Pouw & Dixon, 2019b among others). Thus, the current study is the first to our knowledge to show empirically that while non-referential gestures may be particularly suited to be produced in a rhythmic fashion, non-referential RGGs are not more rhythmic or isochronous than referential ones.

The final objective of this study was to assess the relationship between rhythmic gesture production and speech prosody. We specifically wanted to investigate the correspondence between rhythmic gesture apexes and APs, and to explore whether the rhythmic production of gesture could be related to speech prosody

in terms of accentuation or phrasing at the level of the AP. Regarding the correspondence between APs and RGG apexes, we found that most APs contain only one RGG apex. Interestingly, only eight out of 708 APs in our dataset received more than two apexes. This finding is particularly interesting as it seems to suggest that these rhythmic gestures are working within the AP in a very similar fashion to pitch accentuation, where the AP can only receive a maximum of two pitch accents – an optional IA and an obligatory FA on the last full syllable in the phrase (see **subsection 3.1.2**). However, the tendency of RGG apexes to have a one-to-one relationship with subsequent APs within the RGG did not prove to be the majority. In other words, subsequent APs within an RGG were often skipped or doubly marked with gesture. Importantly, these patterns did not seem to be driven by pitch accentuation patterns. That is, APs did not tend to be doubly marked by gesture because they were realized with two pitch accents. The findings are in line with those from previous studies on English which have suggested that RGG apexes tend to be produced largely independent of pitch accentuation (McClave 1994; Loehr, 2007). Crucially, the results of the current study show that the relationship between prosodic structure and the rhythmic production of gesture is not entirely independent, at least in French. The results of the linear regression model indeed suggest that the time span between AP onsets significantly predicts the time span between RGG apexes. Thus, as APs become longer, as do the distances between subsequent apexes.

Thus, there does seem to be a complex rhythmic relationship between APs and RGGs, which needs further clarification. Researchers have suggested that rhythm is a multi-level phenomenon, where it can be instantiated at various levels (e.g., Pouw et al., 2021). In terms of speech, it can be instantiated at the levels of the syllable, the foot, the prosodic phrase, or even between pauses (Couper-Kuhlen, 1993; as cited by Astésano, 2022). Similarly, gestures can subsequently occur within a clause to mark multiple stressed syllables, at the clausal level marking syntactic conventions (e.g., to reflect verbal conventions for describing motion, Kita & Özyürek, 2003), or show similar gesture forms across multiple minutes for discourse cohesion (McNeill's "catchments", 1992). More recent studies have approached gesture-speech synchrony from the perspective of the Dynamic Systems Theory (McNeill, 1992, 2005; Iverson & Thelen, 1999; Rusiewicz et al., 2014; Pouw & Dixon, 2019a), which assumes that rhythmic behavior is oscillatory in nature, and multiple oscillators may couple or entrain (i.e., they may influence each other) "resulting in either an identical rhythmic pattern or a compromise rhythmic pattern somewhere in between the two patterns relative to when they are produced in isolation" (Rusiewicz et al., 2014, p. 284, see also O'Dell & Nieminen, 1999 in regards to speech rhythm; Iverson & Thelen, 1999 in regard to speech-gesture association, particularly in development).

Similar to observations by McClave (1994, p. 56) and Loehr (2007), there were many cases observed where the gestural rhythm would

either double or halve, suggesting some sort of mathematical periodicity. In the current dataset, such phenomena led to certain APs in the rhythmic group to be skipped by gestural marking during the RGG, and others to receive two RGG apexes. A preliminary study by Pouw & Dixon (2019a) used cross-wavelet analysis to inspect the shared periodicities between gesture (measured in terms of velocity of movement) and speech (in terms of the amplitude envelope) and found that shared periodicities are statistically reliable at slower timescale (specifically 2-6s, or the timescale of the sentence or clause). The timescale of the RGGs as defined in the current study tended to be much shorter (around 1s), thus one potential explanation for these variances in the rhythmic production of gestures could be that gesture oscillator regularly decouples from a speech oscillator (as measured by the amplitude envelope from Pouw & Dixon, 2019a). However, it still remains unclear whether the gesture oscillator could then be coupling with other "speech oscillators", for example, marking syllabic rhythm. For example, **Figure 3.9** (upper panel) shows the annotations of an RGG co-produced with the utterance "*J'ai pu observer que le niveau d'énergie de la personne montait*." ("I could observe that the person's energy level increased"). The lower panel shows a rough schematic representation of how the gesture oscillator may entrain at various levels, where the first two apexes of the rhythmic group co-occur on each syllable within the AP (*j'ai pu*, AP 220), potentially indicating that gestural rhythm is being instantiated at the syllabic level, followed by a single apex in AP 222 and another

at AP 224, potentially indicating that gestural rhythm is being instantiated at a higher phrasal level.

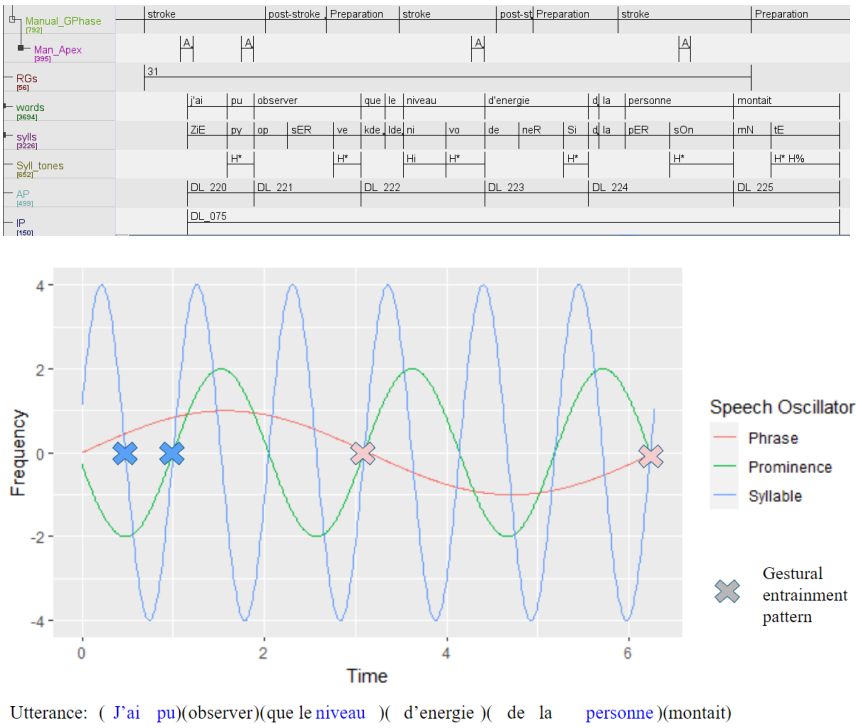

Figure 3.9 Example of multilevel rhythmic entrainment. **Upper panel:** Annotations of an RGG co-produced with the utterance "J'ai pu observer que le niveau d'énergie de la personne montait." (*I could observe that the person's energy level increased*) taken from the French M3D-TED corpus by speaker DL (TEDx Talks, 2016) at 08:23. **Lower panel:** rough schematic representation of speech oscillators and gestural entrainment patterns, where the first two apexes may be marking a syllabic rhythm (blue crosses), and the last two at the phrasal level (red crosses).

The results of the current study along with those of previous studies have suggested that accentual rhythm (accounting for both IA and FA) does not seem to be the guiding factor for gestural rhythm.

However, the results of the current study show how prosodic phrasing still influences the production of rhythmic gestures, which would be in line with Pouw & Dixon (2019a) in that entrainment between gesture and speech oscillators is most stable at slower timescales (in this case, the AP). Future studies will need to further explore this relationship between prosody and gesture, and may thus employ more sophisticated techniques to mathematically measure periodicity, such as cross-wavelet analyses, (Pouw & Dixon, 2019a, 2019b; see also Pouw et al., 2020) or other measures (see Burchardt & Knörnschild, 2020 for an interesting proposal to standardize rhythmic analyses for complex acoustic signals).

The current study gives further evidence that prosodic phrasing plays a role in the temporal alignment between (rhythmic) gesture and speech, suggesting that in French, the syllable prominence alone is not sufficient, but rather syllables combine to form an AP which is then comparable to a foot in Germanic languages, and act as a guiding for the production of rhythm in speech and gesture. This raises questions about English, as there is no evidence of prosodic phrasing lower than the level of the intermediate phrase. What, if any, prosodic factors lower than pitch accent can be affecting this relationship? McClave (1994) mentions that the core attractors of RGG apexes seem to be nuclear pitch accentuation and multisyllabic words with primary stress. Therefore, future studies may look at the metrical strength of non-pitch accented words to determine whether metrical structure acts as a guiding light for these RGGs in English. Conversely, future studies are needed to assess the role of higher levels of prosodic phrasing in the gesture-

speech alignment interface. Perhaps a more stable rhythmic relationship can be found when RGGs occur when, for example, the AP occurs at an intermediate phrase (ip) or intonational phrase (IP) boundary (as suggested by the findings in Pouw & Dixon, 2019a).

All in all, the current study sheds light on gesture-speech temporal alignment patterns in French and how the phrasal positions in the AP are gesture attractors. In terms of rhythm, while not strictly a one-to-one relationship between gesture and prominence has been found, the AP seems to guide the rhythmic timing of gestures, with referentiality having no effect on the isochronic nature of rhythmic gestures groupings.

4

CHAPTER 4: VISUALIZING PROSODIC STRUCTURE – MANUAL GESTURES AS HIGHLIGHTERS OF PROSODIC HEADS AND EDGES IN ENGLISH ACADEMIC DISCOURSES

4.1. Introduction

Gesture and speech prosody are closely temporally coordinated (see Shattuck-Hufnagel & Ren, 2018 for a recent review). Research has shown a tight temporal relationship between prominence-lending tonal movements (i.e., pitch accentuation) and prominence in gesture (i.e., strokes and apexes, or the interval or point in time respectively in which the peak of effort in the gesture occurs). However, prosodic structure consists of not only prosodic heads (e.g., pitch accentuation) but also of prosodic edges (loosely understood as initial and final positions within a prosodic phrase). While initial evidence has suggested that prosodic phrasing indeed plays a role in the temporal execution of gesture (Esteve-Gibert & Prieto, 2013; Loehr, 2012; Krivokapić et al., 2017), to our knowledge, no previous studies have assessed the value of prosodic edges in the attraction of manual gestures by at the same time controlling for the relative degree of prominence associated with the pitch accent in an independent manner. The current study adds to our knowledge of how gestures temporally associate with speech by assessing the following three questions, namely (a) whether the strokes and apexes of manual gestures associate with pitch accented syllables; (b) whether gesture strokes align more with nuclear than prenuclear pitch accents at the intermediate phrase level; and (c) whether this relationship is driven by prominence relations or by phrasal position. A prosodic and gestural analysis of the English M3D-TED corpus was carried out, which contains a total of five academic lectures with over 23 minutes of multimodal speech. Results revealed that while the majority of strokes of manual

gestures (85.99%) overlapped a pitch accented syllable, similar to rates that have been reported before, apex alignment was shown to occur at relatively low rates (50.4%). At the phrasal level, crucially our results also showed that strokes tend to align with phrase-initial prenuclear pitch accents over nuclear accents, and this relationship is not driven by prominence relations. All in all, these findings show that not only prosodic heads, but also prosodic edges (referring to the first prenuclear pitch accent), act as strong attractors of manual gestures, and that future research about gesture-speech temporal association should take this modulating factor into account.

4.1.1. Gesture types, landmarks, and their association with prominence

Speakers naturally make use of a variety of multimodal resources in communication. Two such resources are the use of speech prosody and co-speech gestures. Speakers move their hands and body in a communicative way and evidence from the gesture field has revealed that (a) manual gestures are semantically and pragmatically coherent with speech, and (b) a tight temporal relationship exists between prominence-lending tonal movements (i.e., pitch accentuation) and prominence in gesture (for more information, see McNeill's 1992 synchrony rules, p. 26-29; **subsection 1.1.** of the current thesis). Indeed, initial qualitative observations that the stroke of a manual gesture (that is, the interval in time in which the peak of effort in the gesture occurs, Kendon, 1980; McNeill, 1992) generally co-occurs with or slightly precedes

a stressed syllable have given way to much more quantitative analyses across languages. Since then, numerous studies have investigated the relationship between prominence in speech and prominence in gesture. These studies have varied widely in terms of the type of speech that is studied (i.e, natural, (semi-) spontaneous speech vs. speech produced in laboratory-controlled tasks), the target types of gesture studied, as well as the landmarks that have been chosen in speech and gesture to assess synchrony.

In terms of gesture types, among the most widely used gesture typologies is that proposed by McNeill (1992) which divides manual co-speech gesture into iconic, metaphoric, deictic, and beat gestures. Iconic gestures imagistically represent concrete objects or ideas in speech, while metaphoric gestures imagistically represent abstract ideas in speech. Deictic gestures refer to spatial relations with concrete or abstract entities (i.e., pointing). Finally, beat gestures have been described as gestures which do not represent semantic content in speech, but rather are gestures with simple biphasic movements of the hands that associate with speech prominence and rhythm and have special discourse-pragmatic functions. Such an approach to the classification of beat gestures by their prosodic and pragmatic characteristics (as basically being the gesture types that are associated with prosodic prominence) (a) seems inconsistent with the aforementioned synchrony rules (which applies to all gestures), and (b) has not been tested empirically. Crucially, most studies on the temporal association between speech and gesture prominence have either focused on one type of gesture, or not considered the effects of gesture type at all. The current

approach divides gestures into two broad categories based on their referentiality: gestures which are referential to semantic content in speech (corresponding to McNeill's Iconic, Metaphoric, Deictic types) and those which are non-referential, which do not show any semantic content in speech (englobing McNeill's beat gestures). Given the previous claims by McNeill on beat gestures being associated with prosodic prominence, there is clearly a need to further assess the temporal alignment patterns of referential vs. non-referential gestures.

In terms of gesture landmarks, studies generally assessed the alignment behavior of two positions within the phasing structure of a gesture that represent its prominence, namely the *stroke* of a gesture (the most prominent interval of movement that bears gestural meaning, usually identified by factors such as hand shape, speed, direction, etc.) and the *apex of a gesture,* an instant in time which represents the "kinetic goal of the stroke" (Loehr 2004, see also 2007, p. 190; this phenomenon has alternatively been termed "hits" in some studies, e.g., Yasinnik et al., 2004, see also Rohrer et al., 2021 for more details) and generally refers to points in time where the hands suddenly stop or change direction, corresponding to moments of zero acceleration. Some studies have assessed the temporal association patterns of gestures by analyzing the overlap between gestural strokes (taking into account all gesture types) and prosodic prominence. For example, Karpiński et al. (2009) studied task-oriented dialogues carried out by Polish speakers and found that 75% of gesture strokes overlap with a strong metrical prominence according to the RaP method of annotation (Dilley,

2005). Another study by Shattuck-Hufnagel & Ren (2018) investigated the temporal overlap of gesture strokes and ToBI (Tones and Breaks Indices) defined pitch accented syllables in English academic lectures. The authors found that 83.13% of non-referential strokes overlapped with syllables that were annotated as having a pitch accent, with similar rates for referential gestures (82.85%).

Another commonly studied gestural landmark for testing temporal association is the apex (that is, the point in time in which the movement reaches its kinetic "goal" and can be identified through sudden stops in movement or changes in direction, see Loehr, 2004). A number of laboratory-based studies have found a robust relationship between gesture apexes and pitch accentuation. For example, Leonard & Cummins (2011) investigated non-referential gesture production by a native English speaker with the aid of motion-tracking devices. They used a reading task where the speaker was instructed to read three different passages (each passage was read two times) and to produce three beat gestures on prominent syllables that were chosen beforehand. Based on the 18 data points from the readings (3 texts x 3 gestures x 2 readings), the authors found that gesture apexes tended to be the least variable in terms of their timing with prosodic landmarks (compared to four other kinematic landmarks: onset of movement, peak velocity of extension/retraction, and offset of movement). They also found that the closest prosodic landmark to gesture apexes was the peak of the pitch within the pitch accented syllable. A number of laboratory-based studies have focused on the production of deictic (i.e.,

pointing) gestures, and how this is modulated by the phonetic realization of the target words. For example, Esteve-Gibert & Prieto (2013) investigated the production of deictic gestures in a picture-naming task carried out in Catalan, where participants uttered target words with varying metrical structures in an embedded sentence. They found that the apex of the pointing gesture occurred during the pitch accented syllable, and that relative to syllable onset, the apex showed a stronger correlation with the pitch peaks than other gestural landmarks (i.e., the stroke onset and offset), regardless of the metrical structure of the target word (see also, Rochet-Capellan et al., 2008).

Studies assessing the apex as the gestural landmark in natural speech data, that is, naturally occurring, (semi-) spontaneous speech, have again not distinguished gestures by their referentiality (accounting for all gestures). Loehr (2004) analyzed a total of 2 minutes and 44 seconds of conversational speech from four speakers. The author subsequently found that gesture apexes occurred within 275ms of a pitch accent 74.8% of the time, with the average distance being 17ms before the pitch accent (SD = 341ms). Jannedy & Mendoza-Denton (2005) analyzed 59 seconds of multimodal speech from an audience member at a public congressional town hall meeting and found that 95.7% of apexes co-occur with pitch accentuation. Similarly, Yasinnik et al. (2004) analyzed approximately five minutes of speech from a single speaker giving an academic lecture and found that 90% of hits (apexes) occurred within the boundaries of a pitch accented syllable. Another study by Esposito et al. (2007) analyzed two 4-

minute Italian dialogues by two native speakers (one male and one female) and reported rates of alignment between hits and pitch accented syllables to be 78% and 84% for each speaker respectively. Finally, in a narrative retelling task carried out by native Turkish speakers, Turk (2020) identified different tonal events (F0 minima and maxima) associated with pitch accents as well as phrase accents and boundary tones, and assessed their temporal relationship with the gesture apex. He found that apexes were mainly attracted to pitch accents. Finally, one recent study has taken advantage of advances in modern technology to assess gesture's anchoring points with speech with kinematic measurements. Pouw & Dixon (2019b) measured the gesture productions of four speakers carrying out a narrative-retelling task with the use of motion trackers. Peak F0 values were then extracted from the associated words and the distance between the pitch peak and three kinematic measures from gesture were assessed, namely peak acceleration, peak velocity, and peak deceleration (the latter corresponding closest to the gesture apex). The distributions showed no significant difference by gesture type. Across all gestures, the two landmarks in gesture closest associated with peak pitch were peak velocity (leading peak pitch by an average of 39 ms) and peak deceleration/apex (lagging behind peak pitch, occurring on average 44 ms later). Thus, the results of this cohort of studies suggest a tight temporal association specifically between apexes and pitch accentuation.

Even though the majority of studies have revealed a tight temporal relationship between prominence in gesture and prominence in

speech (e.g., Loehr, 2004; Leonard & Cummins, 2014; Shattuck-Hufnagel & Ren, 2018, among many others), this is not always the case (none of the percentages reported are at ceiling) and some studies have reported conflicting results. McClave (1994) investigated the timing of 50 rhythmic "beat" gestures produced in conversational speech. She found that these rhythmic gestures did not all align with pitch accented syllables, but rather that one of the gestures within the rhythmic group would align with the tone nucleus, and the others would rhythmically span out, falling on both accented and unaccented syllables. A few laboratory studies have also found conflicting results with deictic gesture production in a picture-naming task. In Dutch, De Ruiter (1998, as cited in Esteve-Gibert & Prieto, 2013) found that lexical stress did not influence the production of deictic gestures (though in a second experiment, the author found that contrastive focus acted as an attractor for gesture apexes), and in English, Rusiewicz (2010) found that pointing gesture apexes tended to align with word onset regardless of the metrical structure of the target words, or their contrastive status.

Thus, even though some conflicting results have been reported, the findings up until now generally show a close relationship between prominence in gesture and prominence in speech. As mentioned before, these results generally apply to all gesture types and to our knowledge no studies have assessed the gesture-speech alignment patterns depending on gesture type (i.e., referential vs non-referential gestures). Second, crucially, it is important to note that not all prosodically prominent positions in speech attract gesture, and it is typically a selection of pitch accentual positions that are

most prone to attract gestures. To our knowledge, very few studies have assessed the role of higher-level prosodic structure in the gesture-speech alignment patterns. Even though several authors have claimed that "nuclear pitch accents within the phrase" are the ones that attract more gestures (e.g., Kendon, 1980; McClave, 1994, 1998), very few studies have empirically assessed this issue. An important question is thus whether degrees of phrasal accentual prosodic prominence are the main factor in the attraction of gesture or their position within a prosodic phrase (that is, whether the nuclearity of a pitch accent plays a role in gesture production). In the next subsection we review the studies that have considered phrasal prosodic structure (incorporating pitch accent nuclearity and prosodic phrasing) in the synchrony between gesture and speech prosody

4.1.2. The role of phrasal prosodic structure in gesture production

In the Autosegmental-Metrical (AM) approach to prosodic theory, nuclearity is defined in terms of phrasal position (see., e.g., Ladd 2008, p. 133; **subsection 1.3.1.** of the current thesis). The term "nuclear" designates the last instance of a phenomenon in a phrase. As such, the nuclear pitch accent is the final pitch accent that occurs, either at the ip- level (thus, being ip-nuclear) or at the IP level (being IP-nuclear). Any pitch accent that occurs in the phrase before the nuclear pitch accent is designated as being prenuclear (and unaccented syllables that are uttered after the nuclear pitch accent are designated post-nuclear). Though not explicitly

integrated in this conceptual definition, many authors agree that in English, the nuclear pitch accent is generally (though not always) the most prominent pitch accent in the phrase (e.g., Calhoun, 2010b; Ladd, 2008). McClave (1998) investigated the position of referential gestures in English spontaneous dyadic conversation. The author found that over half of the gestures co-occurred with the nuclear pitch accent. A more recent study by Turk (2020) compared distances between tonal events and apexes and intermediate phrases differing in nuclearity in Turkish speech. That is, they investigated whether gestures occurred more in nuclear ips (the last ip in an IP) than prenuclear ips. They found that apexes were associated more with nuclear ips, and that the time distances between apexes and tones were shorter than in pre- or post-nuclear ips. These results suggest that gesture production is affected by phrasal position at the ip-level.

Importantly, a few studies have investigated the temporal relationship between the onsets and offsets of prosodic phrasal structure and the temporal realization of gesture, with the overall conclusion that the two are temporally associated and share many characteristics. For example, Loehr (2012) investigated the timing of gesture phrases (i.e., strokes and any associated preparations or holds; henceforth *G-Phrase*) and found that G-Phrase onsets occur in close temporal proximity to ip onsets (see also Guellaï et al., 2014, who assessed the important role of both gesture and prosodic phrases in the disambiguation of syntactic structures of NPs). When considering the alignment between prosodic phrases and gesticular phrases, Turk (2020) also controlled for nuclearity of the ips and

found that the majority of G-Phrase onsets were paired with the onset of a prenuclear ip, and the offsets paired equally between prenuclear and nuclear ip offsets. The author suggests that G-Phrases tend to start close to prenuclear ip onsets, span multiple ips, and end at the offset of the nuclear ip. Furthermore, both the temporal location of pitch peaks in rising pitch accents, as well as pointing gesture apexes are sensitive to an upcoming prosodic boundary (Esteve-Gibert & Prieto, 2013), and gesture lengthens both under prominence and at prosodic boundaries (Krivokapić et al., 2017).

At this juncture, we hypothesize that pitch accents in edge positions (in particular, phrase-initial and phrase-final positions) might be able to display strengthening effects and that these positions will attract both prosodic and gestural prominence. From the prosodic angle, there is clear evidence that prosodic structure (and specifically prosodic edges) modulates phonetic realization (see Cho, 2016, for an overview). Up until now, most empirical studies have focused on *boundary-related prosodic strengthening* phenomena (e.g., spatial and temporal expansion of articulation that arises in the vicinity of prosodic edges, especially in association with domain-initial position (also known as *domain initial strengthening*). While pre-final lengthening seems to privilege domain-final strengthening (where lengthening is larger at IP- and ip-final positions than medial and initial positions), other phenomena seem to display *initial strengthening effects*. Bolinger (1985) described how some pitch accents that are located at prosodic edges have phrasal marking effects. He specifically offers

the example of reciting a list (e.g., “One, two three, four, five,” p. 85), whereby the first and last element in the list receive a pitch accent. He described how the last item in the list, by default, would receive a nuclear pitch accent, yet an “attention-getting” accent may occur towards the beginning of the phrase. In terms of intonation, evidence has been found in French that phrase-initial F0 rises at the smallest phrase level (the AP, or accentual phrase following Jun & Fougeron, 2000, 2002) tend to occur more frequently at IP-initial positions than in IP-medial ones (e.g., Astésano et al., 2007; see also Fougeron & Keating, 1997; Portes et al., 2012). The edge strengthening hypothesis has been further reinforced when looking at cases of “stress shift” (e.g., the shift of the pitch accent in a polysyllabic word such as “MassaCHUsetts” - capital letters indicating pitch accented syllable - from the third syllable to the first syllable when occurring in contexts such as “the MASsachusetts MIracle”). Investigating such cases in a radio news corpus, Shattuck-Hufnagel et al. (1994) found a significantly higher rate of word-initial accentuation within words that carry the first (or only) accent of a phrase, relative to words with phrase-medial or final accents. The authors thus claim that “speakers seek to actively indicate that a new intermediate intonational phrase has begun by placing a pitch accent on the first accentable syllable” (p. 382), which also coincides with strategies to avoid pitch accent clash in English.

All in all, to our knowledge very few studies have assessed the role of phrasal prosodic structure in the attraction of association of gestures. Despite the fact that clear evidence exists of domain-initial

effects in the realization of phonetic and prosodic characteristics, to our knowledge only **Chapter 3** has found such an effect for gesture attraction specifically to the left edge of the prosodic phrase. More specifically, to our knowledge no previous studies have assessed the role of pitch accent nuclearity in the attraction of gestures by controlling independently for relative degree of prosodic prominence.

4.1.3. Motivation and research questions

The present study aims to contribute with more evidence to the previous work on the role of complex prosodic structure (e.g., prosodic heads or speech prominence, as well as prosodic edges in terms of accentual positions within the phrase) in the gesture-speech alignment interface. Specifically, it has two objectives, namely (a) to assess the temporal overlap between manual gesture strokes and apexes (both referential and non-referential) with pitch accented syllables; and (b) to assess the role of pitch accent nuclearity (ip-prenuclear vs ip-nuclear) on gesture production, specifically to determine if the location of manual gesture is driven by relative degrees of pitch accentual prominence or by a phrase-initial edge effect. No study to our knowledge has thoroughly investigated the effects of pitch accent nuclearity (prenuclear vs. nuclear) on gesture production within the ip-level in English discourse while controlling for relative degree of prosodic prominence, and position within the phrase. The current study aims to respond to three research questions:

1. Do gesture strokes and apexes align with pitch accented syllables in English TED Talks, and is this relationship modulated by referentiality?
2. Do gestures associate with nuclear pitch accents more than prenuclear pitch accents?
3. Is this relationship driven by relative prominence relationships or phrasal position?

In terms of the first research question, we hypothesize that strokes will be largely aligned with pitch accented syllables (around 80%, as per Shattuck-Hufnagel & Ren, 2018). By contrast, apexes will largely align with pitch accented syllables as well, but this relationship may be more variable than strokes (e.g., Pouw & Dixon, 2019b). In terms of the second question, following recent work on domain initial prosodic strengthening effects, as well as initial results on edge-initial positions acting as gesture attractors (e.g., **Chapter 3** of this thesis) and that gestures tend to begin at the onset of ips (Loehr, 2004, 2012), we expect that prenuclear pitch accents will also be key in the temporal association between prosody and gesture, in the sense that they will attract the realization of gesture. Furthermore, we predict that this relationship will not be directly driven by relative prominence but rather this relationship will be modulated by positional effects in higher-level prosodic structure. Specifically, we expect to find evidence of domain-initial effects, with gestures mainly co-occurring at phrase-initial positions, as these positions have been described as "attention-getting", marking the onset of a new prosodic phrase

(e.g., Bolinger, 1985; Shattuck-Hufnagel et al., 1994). The upcoming **Section 4.2** of the chapter will describe the corpus and annotation procedures, as well as the statistical analyses used to assess the aforementioned hypotheses. **Section 4.3** will present the results, and **Section 4.4** will offer a discussion of the results in the context of the current existing literature on the topic.

4.2. Methods

4.2.1. Materials: The English M3D-TED Corpus

The English M3D-TED corpus was used in the current analysis. The audiovisual corpus contains over 23 minutes of multimodal annotated speech and gesture from five different native adult American English speakers giving a TED Talk (mean duration per speaker: 4m 47s). The corpus contains a total of 1156 gesture strokes, 1307 apexes, and 2033 pitch accented syllables. After removing stretches of silence or disfluent speech, a total of 1139 strokes and 1257 apexes remained in the database for analysis.

TED talks are a form of academic speech that has been described as a “hybrid genre” (Caliendo, 2012, p. 101, as cited in Matiello, 2019). Similar in format to a conference talk (a presentation with a limited time slot, given by an expert), TED talks differ in that the members of the audience are often not specialists in the field. This results in a rather informal register being adopted by TED speakers which is more similar to spontaneous conversation, and of particular interest is the use of narration within the genre (Mattiello, 2017; see

Mattiello, 2019 for an overview). Such contexts make TED Talks an ideal genre for the study of gesture, as TED speakers are generally quite expressive and a good number of gestures typically appear in TED Talks (see, e.g., Harrison, 2021). Specifically in the English TED Talk corpus, the mean rate of words per manual gesture, considering both referential and non-referential, is 3.93 (i.e., a gesture is produced approximately every four words on average). Regarding the naturality of the data, though TED talks are oftentimes rehearsed and/or trained, the official TED guide to public speaking (Anderson, 2016) does not give details on how speakers should employ specific prosodic or gestural features in their speech. In fact, the guide proposes that speakers should speak naturally and conversationally. Specific points regarding the use of prosody include tips such as to use varied prosody (speech rhythm, intonational patterns, etc.) that match the meaning to be conveyed. Similarly, in terms of gesture, the guide proposes that speakers move intentionally and make use of their hands and arms to amplify their message in speech. In all cases, however, the guide highlights that this should come naturally and that there are no "rules" for speakers to follow. Thus, TED Talks can be classified as natural, academic style discourse.

4.2.2. Data annotation

The English M3D-TED corpus was independently annotated for prosody and gesture. The entire corpus is available online[18] in the

[18] https://osf.io/ankdx/

format of ELAN files (Wittenburg et al., 2006), as well as the M3D labeling manual which explicitly describes the annotation procedure and each tier that is available in the corpus (Rohrer et al., 2021). The following subsections will describe the annotation tiers that are related to the current study.

4.2.3. Gestural annotation

Gestural annotation was carried out by the author of this thesis and a research assistant within the context of developing the MultiModal MultiDimensional labeling system, following the annotation guidelines are fully described in the labeling manual (Rohrer et al., 2021; see also **Chapter 2**). It makes use of the same annotated tiers as described in **Chapter 3** of this thesis, namely the gesture phasing tier and the gesture referentiality tierset.

4.2.3.1. Gesture phasing

Specifically, only manual co-speech gestures were annotated (i.e., meaningful manual movements that act as an utterance, or part of an utterance, as per Kendon, 2004). All gesture annotation was carried out using frame-by-frame analysis in ELAN, with gesture phasing and annotation being carried out without audio. Stroke identification was largely based on the kinematic properties of the movement (salient movements based on speed, hand configuration, etc.). The apex refers to any sudden stops, changes in direction or moments of zero velocity, which can be seen as the peak effort in the stroke (see, e.g., Loehr, 2004; Yasinnik et al., 2004). The apex

was identified in frame-by-frame analysis as corresponding to the frame in which the image of the hand(s) go from blurry to suddenly clear (refer to **Figure 1.2 in Chapter 1**), or the frame immediately preceding a change in the direction of movement. **Figure 4.1** shows an example of the gestural annotation of a G-Unit containing three non-referential gesture strokes

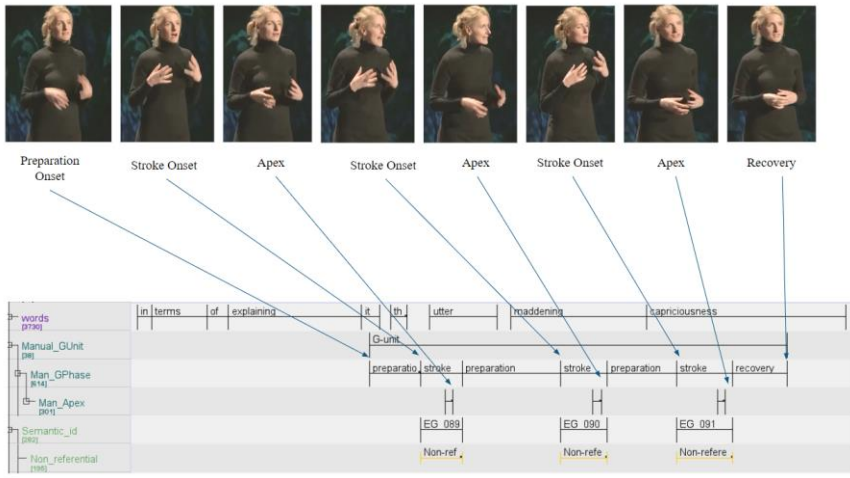

Figure 4.1: Gestural annotation in ELAN of a G-Unit containing three non-referential strokes. Taken from the English M3D-TED corpus by speaker EG (Gilbert, 2009) at 09:53.

Inter-annotator reliability was assessed for each of the two key aspects of gesture phasing, namely gesture phases (i.e., the segmentation of gestures into gesture phases including preparation, stroke, hold, recovery, etc.), as well as apex annotation. The built-in inter-annotator reliability tool in ELAN was used to assess reliability for gesture phasing, which uses an algorithm to assess both temporal overlap as well as value assigned together (Holle & Rein, 2015). The algorithm returned kappa values above 0.76 for

the identification of each type of phase, indicating substantial reliability. Apex location was assessed in terms of distance (in frames) between the two raters and found that 50% of apex annotations were within one frame of each other (33 ms), and 73.8% were within two frames (66 ms). This qualitative assessment of apex coding seems to indicate quite high rates of agreement, particularly considering that Loehr (2004) considered up to six frames of distance as acceptable for agreement.

4.2.3.2. Gesture referentiality

Once gesture phase structure was coded in ELAN without the audio, gesture referentiality (i.e., referential vs. non-referential) was then assessed with the audio. The former have a clear referent in speech through representation (degrees of iconicity or metaphoricity) or by showing spatial relationships (deixis), while the latter do not have a clear referent in speech (e.g., McNeill's "beat" gesture).

Inter-rater reliability for gesture referentiality was assessed using Gwet's Agreement Coefficient 1 (AC1, Gwet, 2008) with MASI distances as the distance metric (Passonneau, 2006; Artstein & Poesio, 2008). The resulting coefficient (which can be interpreted similarly to traditional Kappa) indicated excellent agreement ((AC1 = .895, CI (.856, .933), $p < .001$).

4.2.4. Prosodic annotation

Prosodic annotations were carried out by the author of the present thesis. An orthographic transcription of speech was initially carried out in Praat (Boersma & Weenink, 2022). The transcription was then automatically aligned and segmented into words, syllables, and phones with the Montreal Forced Aligner (McAuliffe et al., 2017).

4.2.4.1. Phonological analysis with MAE-ToBI: Pitch accentuation, pitch accent type, ip and IP boundaries

Prosodic labeling was carried out following the Mainstream American English (MAE) ToBI (Tones and Breaks Indices) system (Silverman et al., 1992; Veilleux et al., 2006). Two main domains were labeled, namely phrasing and pitch accentuation. Regarding phrasing, a breaks tier was used to assess phrasing across four levels, where a 3-break indicates an ip (intermediate phrase) boundary, and a 4-break indicates an IP (intonational phrase) boundary. IPs generally corresponded to entire clauses and generally showed greater pre-final lengthening, often followed by a large pause. Intermediate phrases were identified as smaller groupings of words within the IP, which generally showed some degree of pre-final lengthening or a much smaller pause. Regarding pitch accentuation, a tones tier was used to assign the tonal target to prominent (pitch accented) syllables, as well as phrasal accents (at ip boundaries) and boundary tones (at IP boundaries). Pitch accented syllables are perceived as prominent based on a number of phonetic correlates, usually movements in pitch (i.e., an F0 tonal

target), along with increased duration and intensity. The inventory of pitch accents for MAE ToBI include two simple tonal targets (L* and H*) as well as four complex tonal targets (!H*, L*+H, L+H*, H+!H*) (see, e.g., Veilleux et al., 2006).

4.2.4.2. Annotation of accentual degree of prominence

An additional tier was added in Praat to the ToBI tiers to assess the degree of prominence of each syllable within an IP. Prominence annotation was adapted from the "prominence layer" (tier) described in the DIMA (*Deutsche Intonation, Modellierung und Annotation*) system for German (Kügler et al., 2015). The degree of prominence was annotated for each syllable on a 4-point scale. Syllables with no prominence were encoded as 0. A prominence value of 1 was assigned for weak prominences which do not necessarily coincide with an F0 movement. Such prominences often corresponded to rhythmically-motivated prominences (Calhoun, 2010a), post-focal prominences produced in a reduced pitch register, syllables that contained phrasal accents, or syllables that contain lexical stress. A prominence value of 2 was assigned to strong prominences that coincided with an *f0* tonal movement. Such prominences are said to occur with a typical pitch accent (regardless of its position within the phrase). A prominence value of 3 was assigned to extra strong prominences. These prominences show an additional emphasis that goes beyond a typical 2 prominence, oftentimes showing phonetic differences (e.g., a stronger F0 excursion, greater intensity, etc.) but are phonologically the same as a typical 2 prominence. To carry out the prominence annotations,

the first author listened to the entire IP to identify the most prominent syllables, assigning them 2 or 3 values of prominence. Weaker prominences were then assessed relative to the stronger prominences, accounting for rhythmic constraints (i.e., rhythmically derived full pitch accents or lexical stress). Finally, remaining syllables that were not deemed prominent were assigned a value of 0. **Figure 4.2** shows an example of the prosodic annotations of the sentence "in terms of explaining it, the utter maddening capriciousness" in Praat, where the first tier corresponds to the orthographic transcription (words), the second tier corresponds to the annotations of relative prominence of each syllable ("prom"), and the final two tiers refer to the ToBI annotation (tones, breaks).

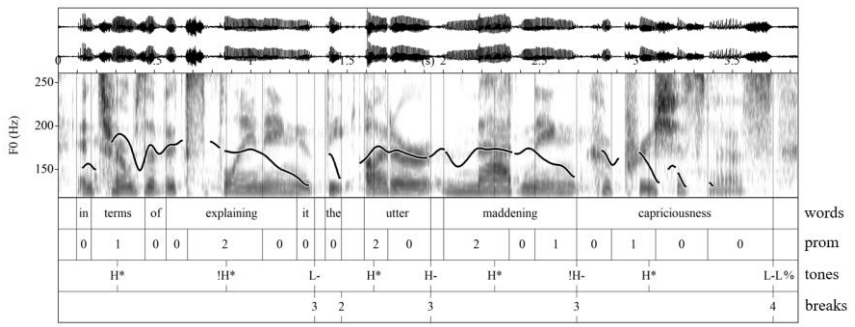

Figure 4.2: Prosodic annotation of an Intonational phrase, composed of four intermediate phrases. Taken from the English M3D-TED corpus, speaker EG at 09:53 (Gilbert, 2009).

Once the prosodic annotations were completed in Praat, the annotations were imported into ELAN. Once in ELAN, two additional tiers were created to facilitate analysis: an intermediate phrase (ip) interval tier and Intonational phrase (IP) interval tier

were created on the basis of the breaks annotations in Praat. The gestural and prosodic annotation data was then exported together in a time-aligned database for further processing in R (R core team, 2021). Finally, two important data transformations were done in R. First, pitch accented syllables were labeled in R as being either prenuclear or nuclear relative to the ip (following the definition that the nuclear pitch accent is the final pitch accent in an ip). Additionally, to operationalize the *relative degree of prominence at the level of the ip* in R (that is, to see which syllables were the most prominent in the phase), each pitch accented syllable was assessed. Specifically, if the pitch accented syllable received the highest prominence value and no other syllable was annotated at the same level of prominence in the ip, it was labeled as the "strongest prominence in the phrase." If two or more syllables shared the highest prominence value in the ip, it was labeled "equally strongest prominence." Finally, if the prominence value was lower than another syllable in the ip, it was labeled as a "weaker prominence in the phrase." Reliability analyses for all prosodic annotations are pending.

4.2.5. Gesture-speech alignment criteria

In order to assess the temporal association of gestures with speech, the temporal overlap between prosodic and gestural landmarks was assessed. While the prosodic landmark of interest was the temporal span of pitch accented syllables, two key gestural landmarks were assessed: the stroke phase, and the apex. First, each stroke was assessed for whether it overlapped with a pitch accented syllable or

not. In other words, if any part of the stroke annotation temporally occurred within any part of the annotation of a pitch accented syllable, then the stroke was considered to have aligned with a pitch accented syllable (e.g., Shattuck-Hufnagel & Ren, 2018). Apexes were also assessed for whether they aligned with pitch accented syllables (i.e., if the point in time that refers to the apex fell within the boundaries of a pitch accented syllable, it was considered as aligned).

4.2.6. Statistical analyses

A series of Generalized Linear Mixed Effects Models (GLMM) were run using the *lme4 package* (Bates et al., 2015) in R. The random effects structure of each model was determined using the *buildmer* function (Voeten, 2022), which compares all potential combinations of random effects and returns the best fitting model. Models which raised convergence issues or overfit the data were re-run as Generalized Linear Models (GLMs). Omnibus test results were then carried out to assess significant main effects, which were then assessed with a series of Bonferroni pairwise tests carried out with the *emmeans* package (Lenth, 2022).

For the assessment of stroke alignment, the GLM with a poisson regression was run with the number of gesture strokes as the dependent variable and included a fixed factor of Gesture referentiality (2 levels: Referential and Non-referential), a fixed factor of Alignment (2 levels: Aligned and Not aligned), and their two-way interaction. The model was offset by the total number of

gestures by type[19]. For the assessment of apex alignment, a GLM with a poisson regression was run with the number of gesture apexes as the dependent variable and included a fixed factor of Gesture referentiality, a fixed factor of Alignment (2 levels: Aligned and Not aligned), and their two-way interaction. The model was offset by the total number of apexes by type[20].

For the assessment of the role of phrasal position (prenuclear vs nuclear pitch accents) in the attraction of gesture, a GLMM with a poisson regression was run with the number of gesture strokes as a dependent variable and a Fixed Factor of Position (2 levels: Prenuclear and Nuclear), with random slopes by speaker[21].

For the assessment of the relative degree of prominence (i.e., whether a syllable was the strongest prominence in a prosodic phrase, shared the strongest prominence with another syllable, or was a weaker prominence), an GLM with a poisson regression was run with the number of gesture strokes as a dependent variable and a fixed factor of Relative Prominence (3 levels: stronger, equal, or weaker prominence), a fixed factor of Gesture Referentiality, and their two-way interaction. The model was offset by the total number of gestures by referentiality[22].

[19] glm(data = df, N_Gestures ~ GestureReferentiality * Alignment, offset=log(Total), family="poisson")
[20] glm(data = df, N_apexes ~ GestureReferentiality * Alignment, offset = log(Total), family="poisson")
[21] glmer(data = df, N_strokes ~ Position + (1 | Speaker), family="poisson")
[22] glm(data = df, N_strokes ~ RelativeProminence*GestureReferentiality, offset = log(Total), family="poisson")

Finally, for the assessment of ip-initial edge effects, a GLMM with a poisson regression was run. The dependent variable was the number of gesture strokes, and included a fixed factor of Position (2 levels: Left Edge and Phrase-Medial). A random effects structure included random slopes and intercepts by speakers[23].

4.3. Results

4.3.1. Temporal alignment between manual gesture strokes and apexes with pitch accented syllables

In response to the first research question, one goal of the current study was to assess the temporal alignment between pitch accented syllables and two gesture landmarks: the stroke and the apex. **Table 4.1** below shows the by-speaker comparisons for both levels of temporal alignment. Though there are some minor differences by speaker, the average rate of alignment between strokes and pitch accented syllables was shown to be 84.32% (SD: 5.71%). Apexes showed substantially lower rates of alignment, with the apex occurring within the pitch accented syllable at an average rate of 49.12% (SD: 7.72%). Gesture referentiality was assessed to determine if potentially one type of gesture (referential or non-referential) showed greater rates of alignment over the other. We found that for strokes, referential gestures aligned with pitch accents 88.34% of the time, and non-referential gestures aligned with pitch accented syllables 82.62% of the time. The GLM model revealed a

[23] glmer(data = df, N_strokes ~ Position + (1 | Speaker), family="poisson")

significant main effect of Alignment ($\chi^2(1) = 615.54$, $p < .001$), indicating that there were more gestures that aligned with a pitch accent than those that did not ($z = -19.395$, $p < .001$), as well as a significant interaction between Gesture Referentiality and Alignment ($\chi^2(1) = 13.01$, $p < .001$). The post-hoc pairwise analyses showed that when gesture strokes aligned a pitch accented syllable, they were equally likely to be referential or non-referential in nature. However, when a gesture did not align with a pitch accented syllable, they were significantly more likely to be non-referential in nature than referential ($z = 3.190$, $p = .006$).

In terms of the apex, non-referential gesture apexes fell within the bounds of a pitch accented syllable 50.91% of the time, while referential gesture apexes fell within the bounds of a pitch accented syllable 47.52% of the time. The GLM model revealed no significant effect of Gesture Referentiality ($\chi^2(1) = 0$ $p= 1$), Alignment ($\chi^2(1) = 0.096$ $p= .756$) nor a significant interaction between Gesture Referentiality and Alignment ($\chi^2(1) = 1.447$ $p= .229$). Taken together, these results indicate no tendency for gesture apexes to be either aligned or misaligned with pitch accented syllables, regardless of gesture referentiality.

The following subsections will first assess whether gesture-speech temporal association is modulated in terms of the nuclear status of the pitch accent (prenuclear vs. nuclear; **Subsection 4.3.2**), and whether this relationship is driven by the relative degree of prominence of the pitch accent and phrasal positioning (i.e., left edge; **Subsections 4.3.3 and 4.3.4.**). Given the low rates of

alignment between apexes and pitch accented syllables, in all subsequent analyses the stroke was chosen as the basic gestural unit of analysis. The low rate of overlap between apexes and pitch accented syllables will be further discussed in **subsection 4.4** below.

Speaker	Stroke Alignment (%)	Apex Alignment (%)
AS	84.08%	44.71%
EG	93.62%	57.53%
ES	84.33%	56.61%
MS	78.62%	46.99%
SJ	80.95%	39.76%
OVERALL	85.99%	50.4%

Table 4.1: The alignment rates between gestures and pitch accented syllables in the English M3D-TED corpus, separated by speaker (Column 1), and between gesture strokes and apexes (Columns 2 & 3)

4.3.2. Temporal association between manual gesture strokes and prenuclear and nuclear pitch accentuation

In terms of prosodic phrasing, the intermediate phrase was chosen as the principal unit of analysis to understand the effect of pitch accent nuclearity for two reasons. First in terms of prosodic phrasing, by choosing a smaller phrase, there is less bias in terms of the number of pitch accents in each category. The number of ips which contained more than one prenuclear accent were relatively few (N = 101) compared to the number containing exactly one prenuclear pitch accent (N = 391). The intonational phrase could be too large as it would naturally have many more potential prenuclear anchoring points (yet only one potential IP-nuclear anchoring point). Second, in terms of gesture, the majority of strokes occurred completely within the boundaries of the ip. Thus, by choosing the ip, we can better control for the number of potential anchoring points of individual gestures, providing more insight into the relevance of the prenuclear/nuclear distinction for gesture attraction.

Of the 1139 gesture strokes that were annotated, 216 were removed from the analysis as they crossed an ip-boundary, leading to ambiguity as to whether the gestures associate with the nuclear region of the first phrase or the prenuclear region of the second prosodic phrase (leaving 923 strokes which occurred completely within the bounds of an ip). An additional 267 strokes were omitted as they either did not overlap a pitch accent (N = 171) or they overlapped multiple pitch accents (N = 96). Finally, gestures which occurred in ips that contained only one (nuclear) pitch accent were

removed (N = 282). A total of 325 gestures remained for analysis, as they occurred within the boundaries of one ip which contained at least two potential anchoring positions (one or multiple prenuclear pitch accents and one nuclear pitch accent), and each stroke overlaps with only one of the potential anchoring points.

Figure 4.3 shows the average number of gesture strokes per speaker associating with each phrasal position, when each ip contains multiple potential prosodic anchoring points (i.e., pitch accents). Of the 325 gestures in such contexts, 194 (59.69%) align with a prenuclear accent, and 131 (40.31%) align with the nuclear accent. Taken at face value, these results suggest that gestures are more attracted to prenuclear pitch accents as prosodic anchoring points than to nuclear pitch accents. However, this analysis contains cases where two gestures may occur within a single ip (essentially marking both prenuclear and nuclear pitch accent). Such contexts do not directly contribute to assessing whether gestures have a clear preference for one position over another, thus a further analysis was carried out, removing cases where multiple gestures occurred within a single ip.

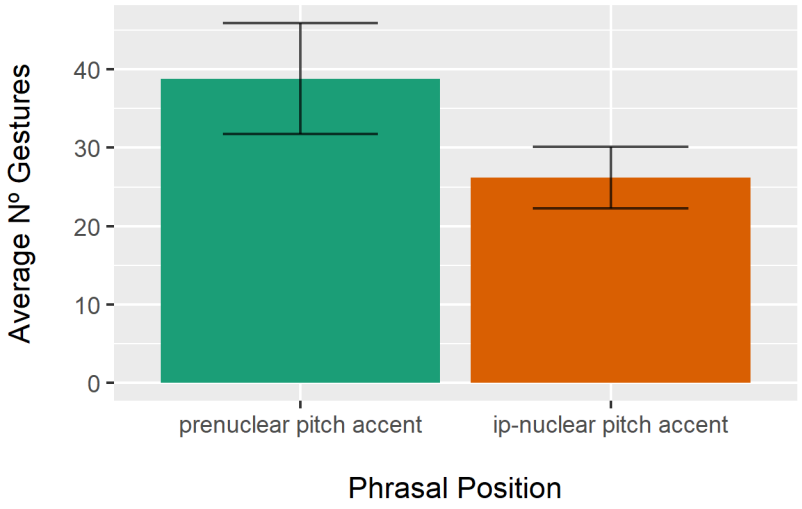

Figure 4.3: The average number of gestures per speaker as a function of the phrasal position with which they align (i.e. prenuclear vs. nuclear) when the ip contains multiple potential anchoring points (error bars show standard error).

The abovementioned finding is further reinforced when inspecting the 119 gestures that occur alone in an ip with only two potential prosodic anchoring points, namely one prenuclear pitch accent and one ip-nuclear pitch accent). In these contexts, the gesture overlaps with the prenuclear pitch accent more often (N = 78, 65.55%) than the nuclear accent (N = 41, 34.45%). **Figure 4.4** shows the average number of gestures aligning with each phrasal position. The results of the GLMM showed a significant main effect of Position ($\chi(1) = 11.2$, $p < .001$), where there were significantly more gestures aligning with prenuclear pitch accents than nuclear ones ($z = 3.34$, $p < .001$).

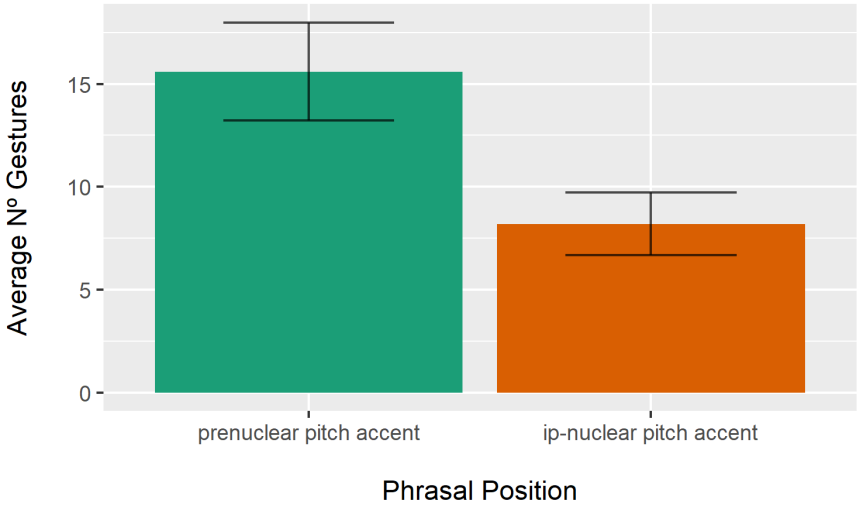

Figure 4.4: The average number of gestures per speaker as a function of the phrasal position with which they align (i.e. prenuclear vs. nuclear) when the ip contains two potential anchoring points and only one gesture (error bars show standard error).

Taken together, these results suggest that prenuclear pitch accents have an important role as anchoring sites for the temporal integration of manual gesture. Specifically, when gestures have multiple potential anchoring points, they tend to associate with prenuclear pitch accents over nuclear pitch accents. The following subsections will assess whether this relationship is driven or modulated by the degree of relative prominence (that is, by assessing whether stronger prenuclear pitch accentual prominences within the ip are attracting more gesture, **subsection 4.3.3**) or a phrase-initial edge effect (a preference for the initial rather than medial positions regardless of their relative prominence, **subsection 4.3.4.**).

4.3.3. Gestural attraction towards prenuclear pitch accents: An effect of relative prominence?

In order to assess whether the attraction of gesture to prenuclear pitch accents is modulated by their relative prominence, we undertook several analyses. First, an initial analysis of the relative prominence ratings across the database showed that nuclear pitch accents were on average perceived to be more prominent than prenuclear pitch accents, with the former receiving an average prominence score of 1.74, and the latter receiving an average score of 1.53.

Figure 4.5 shows the number of gestures that aligned with a pitch accent as a function of the type of pitch accent (prenuclear vs. nuclear) and their relative degree of prominence in the phrase (that is, the pitch accent could be the strongest prominence in the phrase, share the equally strongest prominence with another pitch accent in the phrase, or have a weaker prominence than another pitch accent in the phrase). Of the 325 gestures used for analysis, a total of 194 aligned with a prenuclear pitch accent (the left bar of the **Figure 4.5**), of which 39 (20.1%) occurred in cases where the associated prenuclear pitch accent was the most prominent pitch accent in the phrase, 86 (44.33%) occurred with a prenuclear pitch accent that was assessed as having the same degree of prominence as another pitch accent in the phrase, and 69 (35.57%) occurred in cases where the pitch accent had a relatively weaker prominence to another pitch accent in the phrase. The GLM showed a significant effect of Relative Prominence ($\chi(2) = 9.72$, $p = .008$), but not significant

effect of Gesture Referentiality ($\chi(1) = 2.84$, $p = .092$), nor their interaction ($\chi(2) = 1.4$, $p = .497$). Pairwise comparisons of the significant effect of Relative Prominence showed that when gestures align with prenuclear pitch accents, those accents are significantly more likely to be equally prominent to another pitch accent in the phrase than the strongest one in the phrase. Thus, the attraction to prenuclear pitch accents does not seem to be driven by prominence, and there is no effect of gesture referentiality in this relationship. **Figure 4.6** gives an example of a gesture that associates with a weaker prominence within an ip that contains two pitch accents.

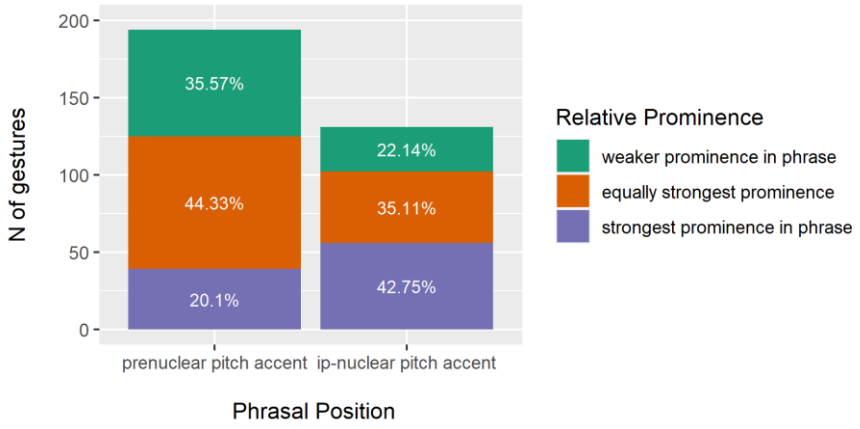

Figure 4.5: The number of gestures per speaker co-occurring as a function of phrasal position (prenuclear vs ip-nuclear) and the relative degree of prominence within the phrase (weaker, equally, or strongest prominence in the phrase, shown as percentage of occurrence).

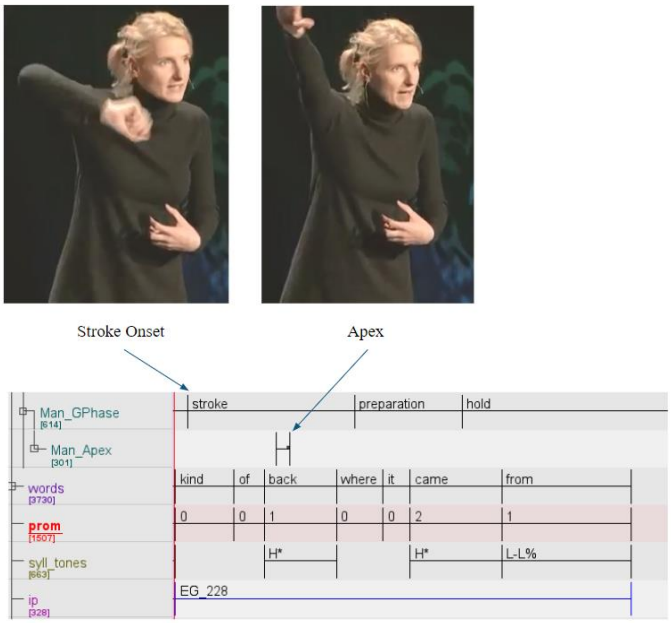

Figure 4.6: Still images of a gesture stroke that associates with a weaker prominence within an ip. The speaker says “kind of back where it came from” with a weaker prenuclear pitch accent and gesture on “back” and a stronger nuclear pitch accent on “came”. Example taken from the English M3D-TED corpus by speaker EG (Gilbert, 2009) at 14:11.

4.3.4. Gestural attraction towards prenuclear pitch accents: A phrase-initial edge strengthening effect?

In the present section we assess the presence of a left edge strengthening effect in the attraction of gesture. We assume that the effects of left edge strengthening should be able to be seen at two levels in the prosodic hierarchy, namely at the level of the ip (where gestures have a clear tendency to mark the first pitch accent in the phrase), and at the level of the IP (where the left edge of IP-initial ips should receive more gestures than IP-medial IPs). Of the 325

gestures that were used in the analysis, 88 were removed as they occurred in phrases where another gesture aligned with an earlier pitch accent in the phrase. Thus, the remaining 237 gestures reflect those that either occurred closest to the left edge (i.e., the first pitch accent in an ip), occurred in a phrase-medial position (where no earlier pitch accent aligned with a gesture), or occurred with the nuclear pitch accent (where no earlier pitch accent aligned with a gesture). **Figure 4.7** shows the distribution of gestures according to their relation to the phrasal edges.

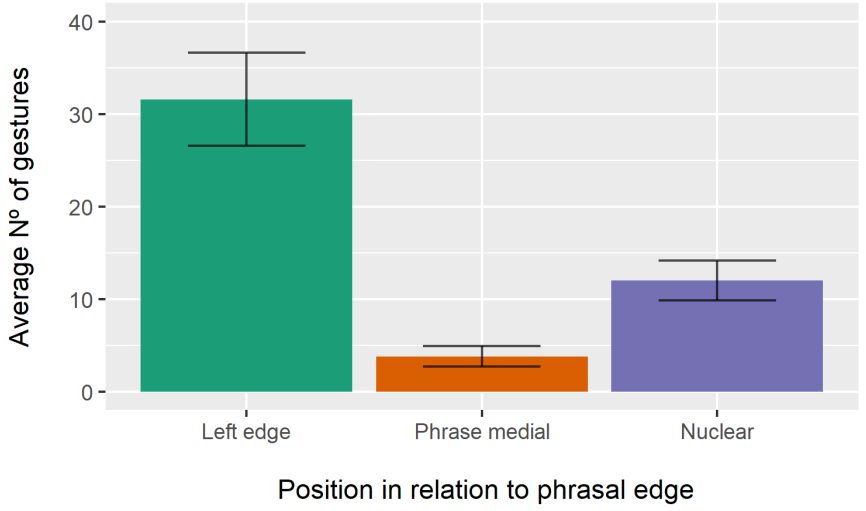

Figure 4.7: The average number of gestures per speaker as a function of its association with prosodic edge position (error bars show standard error).

However, this initial analysis contained ips in which there were only two pitch accents, which would not allow for the assessment of whether pitch accents are going to the left-*most* pitch accent. To refine the analysis, cases where gestures occurred in ips containing only two pitch accents were removed. Of the 53 gestures that

occurred in contexts where the ip contained at least three pitch accents, 23 (43.4%) occurred on the ip-initial pitch accent (marking the left edge), 19 (35.85%) on a medial pitch accent, and 11 (20.75%) on the nuclear pitch accent (see **Figure 4.8**, upper panel). However when taking the average production by speaker (see **Figure 4.8**, lower panel), statistical modeling showed no significant effect of edge marking ($\chi^2(2) = 4.195$, $p = .123$).

At the level of the IP, all gestures which occurred within IPs which contained more than one ip were selected. Of the 117 gestures that marked the left edge, 48 occurred in IP-initial ips, while 69 occurred IP-medially. No further statistical analyses were deemed necessary as the tendency to move towards the left edge of IP-initial ips was not observed.

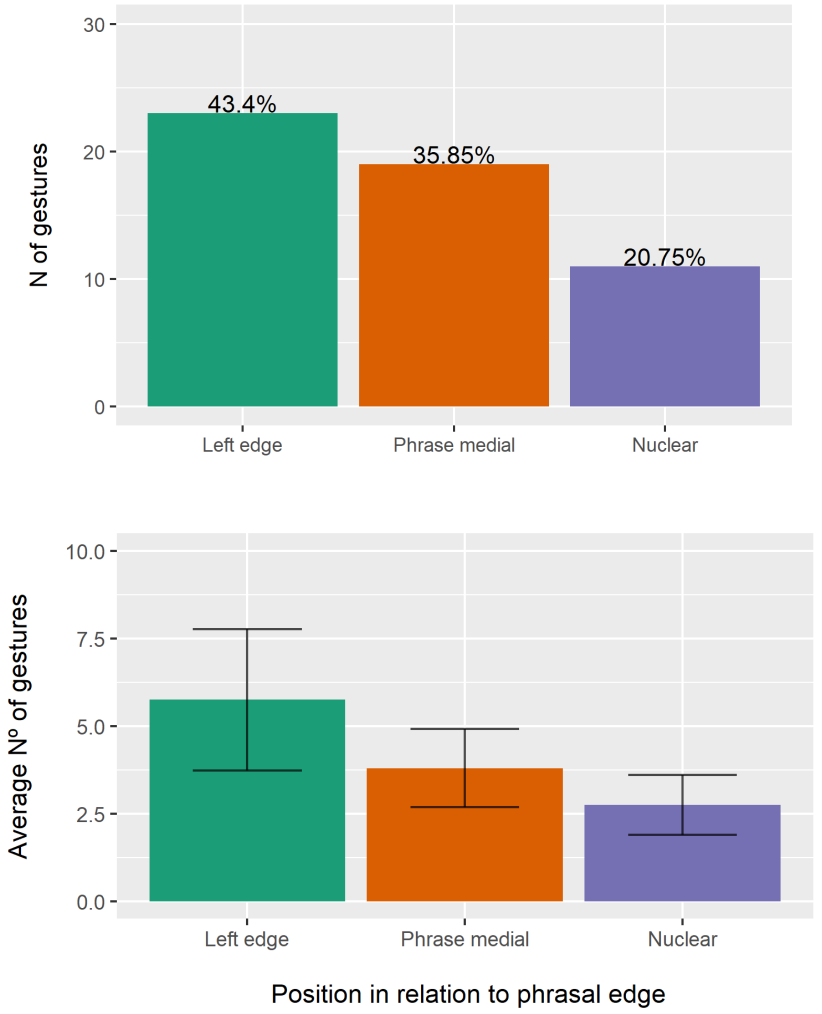

Figure 4.8: Gesture association as a function of phrasal position of the pitch accent. **Upper panel:** Number of gestures across the database occurring in contexts where there are three potential anchoring points by position - left edge (aligning with the first prenuclear pitch accent), phrase medial (aligning with a prenuclear pitch accent that is not the first one in the phrase), and nuclear (associating with the final pitch accent in the phrase). **Lower panel:** The same data as the upper panel, yet showing the average number of gestures produced by speakers (error bars show standard error).

4.4. Discussion and conclusions

The aim of the current study was to investigate the temporal association between manual gesture production and speech by taking into account the role of phrasal prosodic structure in a multimodal corpus of English academic speech (e.g., five TED Talks containing over 28 minutes of multimodal speech). The objectives of the study were twofold, namely (a) to assess the temporal overlap between manual gesture strokes and apexes (both referential and non-referential) with pitch accented syllables; and (b) to assess the role of pitch accent nuclearity (ip-prenuclear vs ip-nuclear) on gesture production, specifically to determine if the location of manual gesture is driven by relative degrees of pitch accentual prominence or by a phrase-initial edge effect. The current study is the first to thoroughly investigate the temporal alignment patterns of strokes/apexes within the boundaries of pitch accented syllables by taking into account the phrase-level constraints in a large English speech corpus.

Regarding the first objective, we found that gesture strokes tended to overlap with pitch accented syllables at an average rate of 84.32%, with no significant differences between referential and non-referential gestures. These results reinforce the idea that all gestures, regardless of their referential nature, associate with prosodic prominence to similar degrees. This finding is in line with what has been reported previously in the literature (e.g., Shattuck-Hufnagel & Ren, 2018 for English; Karpiński et al., 2009 for Polish). Furthermore, it closely resembles previous results that have

compared gesture types, where referential and non-referential gestures have been reported to align at rates of 82.85% and 83.13%, respectively (Shattuck-Hufnagel & Ren, 2018).

However, the results for apex alignment showed an average alignment rate of 50.91% (see also, **Chapter 2** of this thesis). At first, this result may seem surprising given the previous literature on the subject. However, a closer look at the data collection methodologies and alignment criteria adopted in previous studies highlight major differences that allow us to better understand this apparent discrepancy. First, studies that reported high rates of alignment between apexes and pitch accented syllables were mostly laboratory-based studies with controlled conditions, where participants were often instructed to produce gestures on particular words (e.g., Leonard & Cummins, 2011; Esteve-Gibert & Prieto, 2013). Greater variability has been found for studies using (semi-) spontaneous speech, which have tended to use a laxer set of criteria to assess alignment. Second, methodological differences arise between studies in the procedure used to assess temporal alignment. Some studies chose a specific time window to assess alignment. For example, Loehr's (2004) alignment criteria was based on a 275ms time window around the occurrence of a pitch accent. This number was calculated based on his own data, where he calculated the average distance between any gestural events (i.e., begin time and end time of gesture phases, apexes) and tonal events (i.e., pitch accents, phrase accents, and boundary tones), and found that "the majority of the tones ... regardless of type, tended to occur within a distance of 272 msec from the nearest gestural annotation." (p. 103)

Based on this calculation, the author considered any apex occurring within 275 ms of a pitch accent to be “aligned”, as this roughly corresponds to the average word length in his data. An alternative approach employed by Turk (2020) involved a two-step procedure in which first, apexes were paired with the nearest F0 tonal event that shared the semantic meaning as the gesture, and then tested synchronization if it were within the time window of the average syllable duration of the entire corpus (160 ms). Only two studies have specifically assessed the co-occurrence of the apex within the boundaries of the pitch accented syllables. The first study by Esposito et al. (2007) included hits/apexes from other articulators, including the head, shoulders, and eyebrows. When focusing only on the manual gestures, the male speaker produced 138 manual hits (i.e., apexes), of which 87 aligned with a pitch accented syllable (63%) while the female speaker produced three manual hits, of which two aligned with a pitch accented syllable (66.7%). In the second study, Yasinnik et al. (1994) found that in polysyllabic words which contained a hit, 90% also contained a pitch accent (in other words, the authors took a larger time window, the word instead of the syllable, to assess alignment). In monosyllabic words, the authors found rates of alignment closer to those currently reported (65%). However, the authors noted that the results may have shown bias, as the annotator listened to the audio while annotating hits (see Krahmer & Swerts, 2007 for an assessment on how the visual modality may influence the perception of prominence and vice versa).

Why is it that gesture apexes tend to be less aligned with pitch accented syllables than gesture strokes? A closer look at the misaligned apexes in our data showed that these points still co-occurred very close to pitch accented syllables, with over 66% occurring on the syllable immediately preceding or following a pitch accented syllable (with a slight preference for occurring on the syllable following the pitch accented one, and over 91% occurring within a two-syllable distance). Even using different alignment criteria (as previously described), similar results are reflected in Loehr (2004), where the author presents a histogram of the distances (in ms) between apexes and pitch accents. He reports a mean distance of 17 ms, with a standard deviation of 341 ms, indicating quite a large spread. The standard deviation is slightly larger than the average word duration in his data (approx. 300 ms). Thus, it is reasonable to conclude that many apexes do not fall within the bounds of a pitch accented syllable but in neighboring syllables.

These results raise methodological issues for studies investigating the temporal alignment of pitch accents and gestural apexes/strokes in (semi-) spontaneous speech. Although the apex always occurs within the stroke of a gesture, the apex itself merely refers to moments of change in direction or zero velocity. Given the lack of strong association between gesture apexes and pitch accented syllables, a more holistic approach to kinematic measurement might be desirable, as the one used in a recent study by Pouw & Dixon (2019b), which has taken advantage of advances in modern technology to assess potential anchoring points of gestural strokes

with peak pitch using kinematic measurements. The authors assessed three kinematic measures from gesture (namely, peak acceleration, peak velocity, and peak deceleration, i.e., the apex) and found that the two landmarks in gesture closest associated with peak pitch were peak velocity of the gestural stroke (leading peak pitch by an average of 39 ms) and peak deceleration of the gestural stroke (i.e., the apex, lagging behind peak pitch, occurring on average 44 ms later).

The result that it is the stroke of the gesture that more strongly aligns with pitch accentuation is in line with results of other studies suggesting that movement phases of gesture in general may be prominence-lending and thus be more prone to align with prosodic prominence. For example, McClave (1998) found preliminary evidence that occasionally, some speakers may tend to speak and gesture so that pitch and manual movements mirror each other (i.e., the right hand rises as pitch rises, and goes down when pitch falls). Similarly, Ambrazaitis et al. (2020) suggested that not only strokes, but any movement phase of a gesture may be prominence-lending. The authors assessed the temporal association of gesture movement phases with Swedish compound words in a spontaneous speech corpus. Swedish compound words contain two lexical stresses, with the primary stress usually associated with the first. Prosodically, the primary lexical stress is marked by a “late fall” (H*+L) followed by a subsequent peak (H) in the secondary stress, which acts to mark sentence-level prominence (and are, consequently, associated with high levels of prominence). The authors found that not only the stroke, but any potential movement phase could align with

prominent syllables. Specifically, the authors found a preference for stressed syllables to co-occur with preparations and gesture strokes, but showed that holds and even retractions could co-occur with stressed syllables. Similarly, a study by Fung & Mok (2018) described a speaker who regularly showed a lag between the apex of their deictic gestures and prominent syllables. The authors suggest that individual speakers may vary their strategies to achieve gesture-speech synchrony, and that the speaker may have been aligning the movement phase of the stroke with the stressed syllable, as opposed to the apex (see also McClave, 1998 for similar observations regarding by-speaker variability for how manual movements correlate with pitch movement, e.g., the hand rising as pitch rises). Thus, it remains unclear exactly the degree to which the apex can be said to be a meaningful and robust gestural anchor compared to other kinematic landmarks within the stroke or even gestural movements outside of the stroke.

Regarding the second objective of the study on the role of the nuclearity of the target pitch accents (e.g., prenuclear or nuclear pitch accents), the current study unexpectedly found that gestures have a tendency to shift towards prenuclear positions, and this was not driven by relative prominence at the phrasal level. The only previous study to our knowledge that has assessed the relationship between nuclear and prenuclear pitch accentuation is McClave, (1998). Even though the study found that gestures tended to associate with nuclear pitch accents, it only assessed referential gestures and did not offer any quantitative analysis. Similar results to the current study have been found in Swedish compound words.

Specifically, Ambrazaitis et al. (2020, see above) found that when only one of the two syllables in Swedish compound words overlapped with a movement phase, it was almost always the first (primary) lexical stress[24]. Importantly, they found that movement phases associated with the first lexical stress even in cases where the second stress was acoustically considered more prominent. The authors thus suggest that the movement phases of gesture (regardless of type) may be marking the primary stress of compounds (regardless of their relative prominence). By doing so, the authors suggest that gesture may be functioning to identify compound words as a single unit (as opposed to being two separate words), potentially aiding in disambiguation and speech processing for the listener (e.g., Guellaï et al., 2014).

Our results put into question a strict view of the one-to-one relationship between gesture prominence and prosodic prominence, as in our data relative prominence cannot predict the pitch accents that attract gesture alignment. Rather, it seems that gesture may have a preference for prenuclear pitch accents, which regularly mark the left edge of the prosodic phrase. Indeed, when separating prenuclear pitch accents according to their position in the phrase (being at the left edge or being phrase-medial), we found that most gestures associated with the left edge of the ip. This finding is in line with Loehr (2004, 2012), who found that gesture phrases (that

[24] The authors describe their results in terms of gestural movement phases without distinguishing between preparations, strokes, or recoveries. However, a closer look at their published data still shows that when one of the two stresses coincided with a stroke, strokes tended to associate more with the initial stress (N=11) than with the secondary stress (N=8), though the sample size may not have enough power for statistical analyses.

is, the entire gestural movements including the stroke as well as any preceding preparation movements of holds) begin in close temporal proximity to intermediate phrase onsets. However, when controlling for when multiple potential anchoring points were present (removing ips with only one prenuclear pitch accent, or IPs with only one ip), we did not find clear evidence of a specific left-edge marking effect.

It is important to take into account that in order to assess the role of complex phrasal prosodic structure we have to perform data selection to avoid ambiguous cases (i.e., removing gesture strokes which occurred within the bounds of an ip, those that did not align with a pitch accented syllable, or those that overlapped with multiple pitch accented syllables, etc.). This led to fewer observations compared to the database as a whole. Furthermore, while the tendencies seem to be shown across the database at the ip-level, these tendencies are lost when considering variation across speakers. As previously mentioned, such by-speaker variability has been described in previous studies, where researchers find speakers may vary how they align their gestures with prosody (e.g., McClave, 1998; Fung & Mok, 2018). Thus, all in all, even though gestures are indeed marking the left edge more often in the present database, more evidence is needed to shed further light on the left-edge marking effect.

In sum, the current study has found that ip-initial pitch accentuation is a central anchoring site in the gesture-speech temporal interface, regardless of gesture type. Further, prenuclear pitch accents are key

players in the gesture-speech temporal interface, and they act as a prosodic anchor for gesture production, also regardless of their relative degree of prominence in the phrase or the referentiality properties of the gesture. These results thus support the view that higher-level prosodic structure (understood as both prominence and phrasing) is actively modulating the gesture-speech temporal interface.

The present study has some limitations. First, it involves the multimodal analysis of TED Talks, which can be seen as a specific genre of discourse under very particular pretexts (rehearsed speech, given under a time limit and in front of a large audience). Though we argue that such speech is semi-spontaneous (in that it is given without the aid of a script or teleprompter, but given from memory) and natural (no explicit instructions are given to the speakers on how to speak or gesture, see **Subsection 4.2.1**), future studies should work to include other types of discourse, such of spontaneous conversation between multiple speakers, narrative speech, etc. Only by investigating gesture-speech alignment in a variety of discourse settings and following similar methodologies can we better understand multimodal human communication. Second, methodologically, the current study also considered strict overlap from independent annotations in gesture and speech. While such an approach avoids perceptual bias, such a methodological approach is limiting in that a distance of only a few milliseconds could be the determining factors of whether a gesture stroke aligns with a pitch accented syllable or not. For example, a gesture that was considered to not be aligned with a pitch accented syllable may

seem to be perceptively aligned, yet the stroke annotation began only milliseconds after the stressed syllable annotation. The use of independent annotations is quite common in the field as studies have shown that listeners are more likely to perceive gesture and speech as co-occurring (especially when the gesture occurs just before a pitch accented syllable), even if the two do not temporally co-occur (e.g., Leonard & Cummins, 2011, see also Rohrer et al., 2019). However, future studies may consider taking advantage of a perceptual assessment of alignment, as well as a more fine-grained continuous analysis of time alignment, in addition to a strict assessment to avoid edge cases such as that described above in order to achieve a more holistic picture. Additionally, the prosodic annotation was carried out by a single annotator, and thus could benefit from having multiple annotators (e.g., through rapid prosodic transcription, or RPT, Cole & Shattuck Hufnagel, 2016). Third, while the current study has controlled for the semantic contribution of gesture (i.e., whether the gesture is referring to propositional content in speech), other pragmatic factors could be at play, such as the structuring of information in discourse. Indeed, the ip constituents may roughly align with constituents in information structure, such as topic, focus, or discourse referents. As proposed in Ambrazaitis et al. (2020), it may be possible that gesture is marking such constituents as a whole, working conjointly with prosody to offer a multimodal marking of relevant information in discourse. Future studies should address the interplay between gesture marking, prosodic marking, and information structure marking in discourse (see **Chapter 5** of the current thesis).

All in all, the current study has shown that regardless of gesture referentiality, gesture strokes are a robust measure for assessing temporal gesture-speech temporal integration, with apexes not being such a robust measure. Moreover, an important positioning effect was uncovered in the data, where prenuclear pitch accents which often mark the left edge of ip prosodic phrases were key anchoring points for gesture association, regardless of their relative degree of prominence. This suggests that gesture does not only visually highlight prosodic heads, but also prosodic edges. More crosslinguistic work could potentially shed further light on how gestures visually represent prosodic structure.

5

CHAPTER 5: THE MULTIMODAL MARKING OF INFORMATION STATUS OF REFERENTS IN ENGLISH ACADEMIC DISCOURSES

5.1. Introduction

When managing the common ground between speaker and addressee in discourse, speakers may use a number of cues which signal the information status of discourse referents (henceforth ISR, e.g. Noun Phrases or Prepositional Phrases which may be new to the discourse, accessible from context, or given; see Krifka, 2008; Götze et al., 2007). In Germanic languages speakers tend to use pitch accentuation to mark new referents, while given referents are often deaccented (e.g., Kügler & Calhoun, 2021 for a review). Some studies have also suggested a close relationship between pitch accent types and the degree of newness in discourse (e.g., Pierrehumbert & Hirschberg, 1990). Similarly, co-speech gestures seem to be produced with accessible and new referents more than given ones (Debreslioska & Gullberg, 2019; 2020b, Im & Baumann, 2020; among many others), and that, similarly, the production of certain gesture types are related to the newness of information in discourse (McNeill, 1992). However, no study to our knowledge has investigated the joint use of pitch accentuation and gesture production as highlighters of new information in discourse, while accounting for both pitch accent type and gesture type. Thus, the objectives of this study are a) to better understand the multimodal (joint prosodic and gestural) cues to the information status of referents; b) to assess the role of pitch accent type as a prosodic cue to information status via pitch accentuation; and c) to assess the role of gesture type (referential vs. non-referential) as a gestural cue to information status.

A corpus analysis was carried out on the English M3D-TED corpus containing over 23 minutes of multimodal discourse across five speakers, which was independently annotated for gesture, prosody, and information status. We found that both gesture and prosody seem to work together to mark information status, particularly that given referents were found to be more deaccented and produced without gesture than accessible or new referents. However, no significant relationship was found for pitch accent type or gesture type. That is, all pitch accents associate with new information equally, and both referential and non-referential gesture types associate with new information equally. Crucially, in prenuclear pitch accented positions, an interaction between gesture and prosody was found for the first time, showing that gestures marked accessible referents significantly more than given or new ones, playing a complementary role with pitch accentuation. In our view, even these results reflect a good degree of integration between pitch accentuation and the production of gesture, which jointly act as multimodal highlighters of information status, they also reveal a more nuanced situation where gesture is not directly dependent on prosodic structure.

5.1.1. Information structure and the information status of discourse referents

Speakers make use of multiple modes of communication in order to successfully create a meaningful message for their interlocutors. Specifically, speakers can use prosody and gesture to add superimposed layers of meaning to a string of uttered words. For

example, the use of rising intonation at the end of an utterance can change a statement to a question, and gesture can add supplemental information to objects described in speech (showing aspects such as size or shape). Prosody and gesture may even interact, allowing the disambiguation of ambiguous sentences (Guellaï et al., 2014). While previous studies have shown how both prosody and gesture can be used pragmatically to mark new information in discourse (see a review of the literature in **subsections 5.1.2** and **5.1.3** below, see also **subsection 1.4.** of the current thesis), the aim of the current study is to assess how these two modes work jointly in the marking of information structure.

In order to have successful communication, speakers must ensure the common ground (i.e., shared knowledge) between speaker and addressee to be adequately maintained and updated as discourse continues. The way in which we package information to maintain common ground is known as information structure (henceforth, IS; Chafe, 1976, as cited in Krifka, 2008). In other words, speakers must regularly alternate between introducing new information and referring back to old information to move discourse forward in a stepwise, coherent manner. The structuring of information may occur across multiple independent levels, namely, focus (vs. background), topic (vs. comment), contrastiveness, and the information status of discourse referents. The current subsection will focus on the latter aspect of IS (see **subsection 1.4.1** for a brief overview of all the levels of IS).

The information status of discourse referents (henceforth, simply ISR) regards whether discourse referents are new to the addressee, accessible to the addressee through context, or given for the addressee as they have been explicitly mentioned previously in discourse (e.g., Götze et al, 2007; Krifka 2008). A discourse referent is any noun phrase (including pronouns) or prepositional phrase that refers to a specific entity in discourse (e.g., individuals, places, times, or events) that can receive anaphoric or cataphoric expressions. Thus, NPs or PPs that do not refer to discourse referents are excluded, for example in the case of idiomatic expressions ("**On the one hand**…"), expletive "it" ("**It** always rains on Sundays."), or "there" in sentences like "**There**'s a fly in my soup.."[25] In terms of ISR, new referents are those that are completely cognitively inactive — they cannot be deduced through context nor easily predicted by the listener. In **Example 5.1**, the discourse referents *mouse* and *on the beach* are cognitively inactive for the listener and difficult to predict, and are thus new. The referent *he* refers back to the mouse which has already been introduced in the discourse (consequently, it is cognitively active) and is thus given.

[5.1] *A mouse was walking on the beach. Then, he finds a seashell.*

However, *a seashell* would be considered an accessible referent because it is cognitively semi-activated through the previous referent *on the beach* —the mere mention of a beach automatically

[25] Examples taken from Götze et al. (2007)

semi-activates any potential entity related to a beach (such as *sandcastles, surf boards, sunscreen*, and even *a seashell*). Thus, even though it is new to the discourse, the referent is relatively accessible to the hearer. The accessibility of a referent can be assessed through some sort of relationship with a previously introduced referent (as in **Example 5.1**), the situational context in general (for example, if someone asks "Will you pass the salt?" while at the dinner table, *salt* would be considered an accessible referent), or even the supposed world knowledge of the hearer (specific referents whose existence is considered to be common knowledge, such as *the Earth*, *the Sun* or *Barcelona*, etc.). More recently, some studies have begun including contrastiveness along with the ISR under a more global term "informativeness" (as described by Baumann et al., 2019; 2021). The idea is to relate ISR to its role in the focus domain (that is, capture both its more "objective" newness along with its pragmatic role in a proposition, the indication of explicit alternative. Thus, the informativeness of a referent not only increases as a function of its newness (i.e., newer referents are more informative than accessible or given ones) but also as a function of its contrastiveness (i.e., referents in contrastive focus/topics are generally considered more informative than ones in narrow or broad focus).

When managing the common ground in terms of discourse referents, speakers may use a number of cues which signal the ISR. For example, the morphosyntactic form of a referent may act as a cue, where new referents are often produced with an indefinite article while accessible referents can be produced with either an

indefinite or a definite article and given referents can be pronominalized (e.g., Clark, 1975, 1977; Gundel, 1996; Prince, 1992). Another strategy speakers may use is prosody, where in Germanic languages pitch accentuation has been said to play a central role (see Kügler & Calhoun, 2020 for a review). Finally, gestures (and particularly non-referential gestures) have been described as special focus markers (e.g., McNeill, 1992; Loehr, 2012; see Debreslioska & Gullberg, in press, for a review). The following two subsections will describe the multimodal marking[26] of ISR, particularly in terms of prosody (**subsection 5.1.2**) and gesture (**subsection 5.1.3**). As we will see, the majority of studies have independently paid attention to either the prosodic strategies or the gestural strategies used to mark information status of discourse referents, and very few studies have jointly the two strategies (e.g., Im & Baumann, 2020).

5.1.2. The prosodic marking of ISR

Numerous studies have investigated the prosodic marking of information structure, largely supporting the idea that in English as well as other Germanic languages, newer information tends to receive greater prominence through pitch accentuation, while given information is usually deaccented (e.g., Halliday, 1967; Chafe, 1974; Brown, 1983; Gussenhoven, 1984; Hirschberg, 1993; Cruttenden, 1997; Hirschberg, 2002; Prince 1981; Ladd, 1980, 2008

[26] Multimodal marking refers to any aspect of multimodal communication (gesture, prosody, gaze, etc.; see e.g., Mondada, 2016; see **Chapter 1** of this thesis) which encodes, marks, cues, or signals the information structure of speech

among many others; see Kügler & Calhoun, 2020 for a review). One study by Pierrehumbert & Hirschberg (1990) aimed to describe how different intonational contours (and specifically, pitch accents) contribute to the interpretation of discourse. The authors suggest a near one-to-one mapping between pitch accent type and ISR. Specifically pitch accents with a high tonal target (H*) indicate new information, while accents with low tonal target (L*) indicate information that is already in the common ground (i.e., given referents). Rising bitonal pitch accents (L*+H, L+H*) are used to mark elements as contrastive, and downstepping patterns (described as bitonal H+L* pitch accents which are no longer included in the official Mainstream American English ToBI inventory of pitch accents today, but rather are labeled as !H*) correspond to accessible information. Similarly in English, these different pitch accent types are said to correlate with degrees of prominence — namely that deaccented or L* pitch accents are the least prominent, and rising bitonal pitch accents are said to be the most prominent (see, e.g., Ladd & Morton, 1997, Cole et al., 2019 for English; for similar results in German, see Baumann & Röhr, 2015). In other words, degree of prominence and type of pitch accent (based on its tonal target, such as H*, L+H*, etc.) are closely related to IS, with the most prominent encoding being reserved for the most "informative" information (i.e., including contrastiveness with ISR, as per Baumann et al., 2021; See **Figure 5.1**).

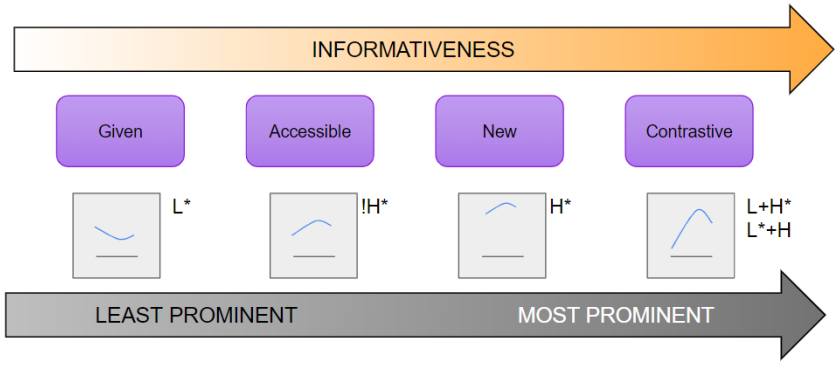

Figure 5.1: Schematic representation of the relationship between prominence, pitch accent type, informativeness (as per Baumann et al., 2019; 2021), and as described in Pierrehumbert & Hirschberg (1990).

However, recent studies in the field of speech prosody have begun calling for a more "probabilistic" interpretation of prosodic categories. That is, there is no one-to-one mapping between pitch accent type and ISR, but rather certain prosodic contours are only *more-or-less likely* to mark certain pragmatic types. For example, Mücke & Grice (2014) showed how the proportion of L+H* pitch accents gradually increased from broad, to narrow, to contrastive focus. Interesting, H* pitch accents were present in all three categories, representing ~20-30% of tokens in each category (see Im et al., 2018 for similar results specifically pertaining to ISR). However, one recent study by Baumann et al. (2021) did not find any categorical or probabilistic relationship with pitch accent type (grouped as no accent, low accent, high accent, rising accent) and informativeness (i.e., ISR + contrast).

In addition to different tonal targets, pitch accents also differ in terms of nuclearity (where according to the Autosegmental-Metrical framework, nuclear pitch accents refer to the last pitch accent in a prosodic phrase and is generally the most prosodically prominent, while prenuclear pitch accents refer to any pitch accents that occur before the nucleus, see e.g., Ladd, 2008; also **subsection 1.4.1.** of the current thesis). While nuclear pitch accentuation has been shown to largely correspond to marking IS, particularly in terms of focus, the stability of prenuclear pitch accent as marking IS has been put into question. For example, Baumann et al. (2021) found that aboutness topics (i.e., referents) in prenuclear position received rising pitch accents, regardless of their information status. Calhoun (2010a) held that prenuclear pitch accents are primarily produced for the rhythmic organization of speech. Similarly, Büring (2007) claimed that such accents are optional and purely "ornamental". However, some studies have shown how the realization of prenuclear pitch accents is modified in terms of ISR. For example, Féry & Kügler (2008) found that both given and new referents in prenuclear positions receive (prenuclear) pitch accents, with given prenuclear referents receiving pitch accents realized with a relatively lower pitch height and smaller range than new prenuclear referents. Similar results have been found in terms of contrastive aboutness topics, prenuclear referents received the same pitch accent type, but contrastive referents received higher and later peaks than their non-contrasted counterparts (Braun, 2006).

Taken together, it seems that there is a probabilistic relationship between prosodic prominence and ISR, where more informative

information (i.e., newer referents) will receive greater prominence which may be encoded by a number of prosodic aspects in different contexts (presence/absence of pitch accent, acoustic cues for those pitch accents, etc.). Now one aspect of speech that is closely coordinated with pitch accentuation is that of gesture. As we will see in the next subsection, in the field of gesture studies, the role of gestures in marking information structure has often been looked at independently of prosody.

5.1.3. The gestural marking of ISR

McNeill (1992) described how gesture production is related to Communicative Dynamism (henceforth, CD). CD was first described by Firbas (1971) as the degree to which an utterance moves the discourse forward. McNeill thus described how speakers tend to use manual co-speech gestures more often when CD is high (i.e., the corresponding information is new or least accessible, thereby pushing communication forward). A number of empirical studies since then have found that gestures tend to mark the introduction of either new referents (e.g., Debreslioska et al., 2013; Debreslioska & Gullberg, 2019; Gullberg, 2003; Levy & Fowler, 2000; Marslen-Wilson et al., 1982; Yoshioka, 2008) or accessible referents in discourse (Debreslioska & Gullberg, 2020b, 2022). Interestingly, Debreslioska et al. (2013) found that gestures may co-occur with given referents particularly when the referents are reintroduced (i.e., they have been explicitly mentioned previously, but reactivated in discourse through the use of a full NP), as opposed to a maintained given referent (i.e., easily accessible to the

hearer and consequently produced with a lexically reduced form such as a pronoun or zero anaphora).

In addition to the production (or non-production) of gestures in order to mark ISR, some studies have shown how ISR (together with the definiteness of the referent) may also guide gestural form and even the semantic content that is encoded in the gesture. For example, one study found that new referents expressed with indefinite nominals or in specialized clause structures that function to introduce entities (e.g., "there was a broom") will lead to more "entity" gestures (e.g., a gesture indicating the shape of the broom), while accessible referents expressed with definite nominals or in less specialized clause structures which merely describe events tended to produce more "action" gestures (e.g., pretending to hold the broom, Debreslioska & Gullberg, 2020b, see also Foraker, 2011). Another recent study by Holler et al. (in press) showed how when a referential gesture is repeated and accompanies a given referent, it is significantly more likely to be produced with a shorter duration, suggesting that "given gestures" are also produced with a reduced form (see also Gerwing & Bavelas, 2004). Finally, subsequent recurrent gestural features (i.e., gestural catchments as per McNeill, 1992) may help build cohesion and aid referent tracking (McNeill & Levy, 1993). However, these previously mentioned studies have mainly focused on referential gestures (i.e., those that illustrate aspects of semantic concepts in speech) or have not distinguished between referential and non-referential gestures (i.e., those which do not portray any semantic concepts in speech).

The theoretical literature on the pragmatic nature of non-referential gestures, particularly in terms of IS marking, offers contradictory views. Namely, non-referential gestures have been described as special markers of focus (Loehr, 2012). McNeill (1992) goes so far to say that one of their functions in narrative speech is to introduce novel characters. However, when McNeill discusses gesture production in terms of CD, he claims that as CD increases, gesture "complexity" increases, namely that "simple" non-referential and deictic gestures occur with information that has lower CD (e.g., pronouns, simple lexical NPs) while "more complex" iconic and metaphoric gestures occur with information that has high CD (e.g., modified lexical NPs, predicates; see also Debreslioska & Gullberg, 2019). The only empirical study to our knowledge to explicitly focus on non-referential gestures is that of Im & Baumann (2020), who investigated the multimodal marking of ISR in terms of both prosody and non-referential gesture in a two-and-a-half minute TED Talk video of an English academic lecture. Although their study was preliminary and thus did not carry out any statistical analyses, their results showed a tendency for non-referential gestures to mark more accessible (separated into "bridging" or "unused" in their study, depending on whether it was available from context or world knowledge) and new referents than given referents. The study also found that over half of the L+H* and H* pitch accents were accompanied by a non-referential gesture, with lower percentages for other pitch accents. However, the authors did not assess interactions between gesture and prosody in terms of the marking of ISR.

All in all, to our knowledge very little is known about how prosodic and gesture prominences are employed jointly by speakers in order to signal ISR. In other words, most studies have only either assessed one mode at a time (i.e., either prosodic or gesture prominence to mark ISR), or have looked at interaction in a piecemeal fashion (i.e., how gesture signals ISR and associates with pitch accentuation). No study to our knowledge has specifically assessed this relationship in a holistic fashion (i.e., the combination of gestural and prosodic cues concurrently to mark ISR).

5.1.4. Motivation and research questions

Previous studies have independently assessed the role of prosodic and gestural marking of ISR. While they have acknowledged a close relationship between prosody and gesture and ISR, no studies to our knowledge have attempted to integrate both components to understand their relative interconnection in the marking of ISR. Thus, the first objective of the current study is to better understand the interconnections between prosody and gesture in the multimodal marking of ISR. Furthermore, previous studies in both prosody and gesture have assumed a one-to-one mapping between multimodal cues, in the sense that gesture association is dependent on prosodic prominence. Yet we have seen that the relationship between the two can be quite complex and dependent on phrasal prosodic structure. There is thus a need to assess the role of ISR in the mapping between prosodic and gestural cues.

If we separate the different components, in terms of prosody, studies suggest that different pitch accent types correspond to different degrees of prominence, which in turn mark ISR (e.g., Pierrehumbert & Hirschberg, 1990). Accordingly, the second objective of the current study is to assess the role of pitch accent type and relative prominence as a prosodic cue to ISR. In gesture, non-referential gestures have been described as special focus markers (e.g., Loehr, 2012); however, studies on the gestural marking of ISR have not explicitly compared the two types of gestures (referential vs. non-referential) as cues to ISR. Importantly, the third chapter of this thesis has shown how in English, prenuclear pitch accents at the ip-level tend to act as strong attractors for gestural production. It could be possible that this relationship is modulated by ISR. Thus, the third objective of this study is to assess the potential interaction between ISR and gesture production, in terms of gesture type, as well as sensitivity to ISR in prenuclear accented conditions at the level of the ip. Thus, the present investigation has three general objectives:

1. How do gesture and pitch accentuation jointly mark ISR in English TED Talks?
2. In terms of prosody, what is the relationship between relative prominence, pitch accent type and ISR?
3. In terms of gesture, does gesture type (i.e., referential vs. non-referential) play a role in marking ISR? Are gestures sensitive to ISR status in prenuclear positions?

First, due to the close relationship between gesture and prosody, we hypothesize that both types of multimodal cues will often act together to mark ISR. Specifically, we predict that both gesture and pitch accentuation will be used to mark new and (to a lesser degree) accessible referents, while given referents will mostly not be multimodally marked. Second, in terms of the correspondence between ISR and pitch accent type, relative prominence, or gesture type, we expect (a) a probabilistic relationship for both pitch accent type and relative prominence, and (b) no significant differences between gesture referentiality types (i.e., referential and non-referential gestures) in the marking of ISR. Finally, we expect fewer new referents in prenuclear positions, causing gestures to associate more with accessible referents in prenuclear positions than given ones. The following section will briefly describe the corpus, annotation procedures, and statistical analyses. **Section 5.3** will describe the results of the study, and **Section 5.4** will offer a discussion of the results to contextualize them with the results from the previous literature.

5.2. Methods

5.2.1. Materials: The English M3D-TED Corpus

The English M3D-TED corpus was used in the current analysis. The audiovisual corpus contains over 23 minutes of multimodal annotated speech and gesture from five different native adult American English speakers giving a TED Talk (mean duration per speaker: 4m 47s). The corpus contains a total of 1156 gesture

strokes, 1307 apexes, 2033 pitch accented syllables, and 1360 referential expressions. After removing stretches of silence or disfluent speech, a total of 1139 strokes and 1257 apexes remained in the database for analysis.

5.2.2. Data annotation

The English M3D-TED corpus was independently annotated for gesture, prosody, and ISR. This same corpus has been previously annotated for prosody and gesture (**Chapter 4** of this thesis), and was annotated for ISR for the current analysis. The entire corpus is available online[27] in the format of ELAN files (Wittenburg et al., 2006), as well as the M3D labeling manual which explicitly describes the annotation procedure and each tier that is available in the corpus (Rohrer et al., 2021). The following subsections will describe the annotation tiers that are related to the current study.

5.2.3. Gestural annotation

Gestural annotation was carried out by the author of the present thesis and a research assistant within the context of developing the M3D labeling system (Rohrer et al., 2021; see **Chapter 2** of this thesis). Only manual co-speech gestures were coded (that is, meaningful manual movements that act as an utterance, or part of an utterance, as per Kendon, 2004). All gesture annotations were carried out using frame-by-frame analysis in ELAN.

[27] https://osf.io/ankdx/

For the present analysis, the following three levels of gesture coding were used, namely gesture phasing, gesture referentiality, and the gestural marking of ISR (i.e., pragmatic domain tier set). For the current study, the pragmatic domain tier was used to annotate the gestures which function pragmatically to mark ISR (see **subsection 5.2.6.** below for the actual procedure used). The following subsections will describe the gesture tiers that are related to the current study.

5.2.3.1. Gesture phasing

Specifically, only manual co-speech gestures were annotated (i.e., meaningful manual movements that act as an utterance, or part of an utterance, as per Kendon, 2004). All gesture annotation was carried out using frame-by-frame analysis in ELAN, with gesture phasing and annotation being carried out without audio. The gesture phasing tier divides the movement into preparation, stroke, retraction, and hold phases. The gesture stroke was identified by the kinematic properties of the movement (salient movements based on speed, changes in handshape, etc.). Importantly, initial passes of the phasing coding were carried out without access to the audio, so as to avoid influence from the speech stream (i.e., for more details, see Rohrer et al., 2021; see also **Chapter 2** of this thesis).

The built-in inter-annotator reliability tool in ELAN was used to assess reliability for gesture phasing, which uses an algorithm to assess both temporal overlap as well as value assigned together (Holle & Rein, 2015). The algorithm returned kappa values above

0.76 for the identification of each type of phase, indicating substantial reliability.

5.2.3.2. Gesture referentiality

Once gesture phase structure was coded in ELAN without the audio, gesture referentiality (i.e., referential vs. non-referential, coded within the semantic tierset) was then assessed with the audio. Referential gestures have a clear referent in speech through representation (degrees of iconicity or metaphoricity) or by showing spatial relationships (deixis), while non-referential gestures do not have a clear referent in speech (e.g., McNeill's "beat" gesture).

Inter-rater reliability for gesture referentiality was assessed using Gwet's Agreement Coefficient 1 (AC1, Gwet, 2008) with MASI distances as the distance metric (Passonneau, 2006; Artstein & Poesio, 2008). The resulting coefficient (which can be interpreted similarly to traditional Kappa) indicated excellent agreement ((AC1 = .895, CI (.856, .933), $p < .001$).

5.2.4. Prosodic annotation

Prosodic annotations were carried out by the author of the current thesis. An orthographic transcription of speech was initially carried out in Praat (Boersma & Weenink, 2022), which was then automatically aligned and segmented into words, syllables, and phones using the Montreal Forced Aligner (McAuliffe et al., 2017).

5.2.4.1. Phonological analysis with MAE-ToBI: Pitch accentuation, pitch accent type, and phrasing

Prosodic labeling was carried out following the Mainstream American English (MAE) ToBI (Tones and Breaks Indices) system (Silverman et al., 1992; Veilleux et al., 2006). Two main domains were labeled, namely phrasing and pitch accentuation. Regarding phrasing, a breaks tier was used to assess phrasing across four levels, where importantly a 3-break indicates an ip (intermediate phrase) boundary, and a 4-break indicates an IP (intonational phrase) boundary. IPs generally corresponded to entire clauses and had greater pre-final lengthening, often followed by a large pause. Intermediate phrases were identified as smaller groupings of words within the IP, which generally showed some degree of pre-final lengthening or a much smaller pause. Regarding pitch accentuation, a tones tier was used to assign the tonal target to prominent (pitch accented) syllables, as well as phrasal accents (at ip boundaries) and boundary tones (at IP boundaries). Pitch accented syllables are perceived as prominent based on a number of phonetic correlates, usually movements in pitch (i.e., an F0 tonal target), along with increased duration and intensity. The inventory of pitch accents for MAE ToBI include two simple tonal targets (L* and H*) as well as four complex tonal targets (!H*, L*+H, L+H*, H+!H*) (see, e.g., Veilleux et al., 2006).

5.2.4.2. Annotation of accentual degree of prominence

As previously reported (see **subsection 5.1.2**), some studies have suggested that the different tonal configurations and phonetic realizations of pitch accents correlate to different degrees of prominence — specifically that higher pitch peaks and greater pitch excursions seem to translate into greater perceived prominence (e.g., Ladd & Morton, 1997), and this conjointly prosodically marks ISR. However, given the more recent probabilistic perspective, it seems necessary to independently assess perceived prominence from pitch accent type. To respond to this issue, an additional tier was added to the ToBI tiers in order to assess the relative degree of prominence of each syllable within an IP on a 4-point scale. The coding was adapted from the "prominence layer" (tier) described in the DIMA (Deutsche Intonation, Modellierung und Annotation) system for German, Kügler et al., 2021). These annotations were carried out independently of the pitch accent status of prominent syllables, where syllables with no prominence were encoded as 0. A prominence value of 1 was assigned for weak prominences, while a prominence value of 2 was assigned to strong prominences[28]. A prominence value of 3 was assigned to extra strong prominences. To carry out the prominence annotations, the first author listened to the entire IP to identify the most prominent syllables, assigning them 2 or 3 values of prominence. Weaker prominences were then assessed relative to the stronger prominences. Finally, remaining

[28] It should be noted that level 2 and 3 prominences generally correspond to pitch accented syllables, whereas syllables with level 1 prominences may or may not be considered pitch accented. Importantly, the tonal status of pitch accented syllables (e.g., L* vs. H*) was not considered when assessing relative prominence.

syllables that were not deemed prominent were assigned a value of 0.

Once the prosodic annotations were completed in Praat, the annotations were imported into ELAN. Once in ELAN, two additional tiers were created to facilitate analysis: an intermediate phrase (ip) interval tier and Intonational phrase (IP) interval tier were created on the basis of the breaks annotations in Praat. The gestural and prosodic annotation data was then exported together in a time-aligned database for further processing in R (R core team, 2021). Finally, two important data transformations were done in R. First, pitch accented syllables were labeled in R as being either prenuclear or nuclear relative to the ip (following the definition that the nuclear pitch accent is the final pitch accent in an ip). Additionally, to operationalize the *relative degree of prominence at the level of the ip* in R (that is, to see which syllables were the most prominent in the phase), each pitch accented syllable was assessed. Specifically, if the pitch accented syllable received the highest prominence value and no other syllable was annotated at the same level of prominence in the ip, it was labeled as the "strongest prominence in the phrase." If two or more syllables shared the highest prominence value in the ip, it was labeled "equally strongest prominence." Finally, if the prominence value was lower than another syllable in the ip, it was labeled as a "weaker prominence in the phrase." Reliability analyses for all prosodic annotations are pending.

5.2.5. ISR annotation

ISR annotation was carried out by the first author and a research assistant according to the methods established in the (M3D) labeling system. Specifically, annotation of the ISR was carried out exclusively on the text, so as to avoid any influence from either prosody (in the audio channel) or gesture (in the visual channel). The coding of ISR was adapted from the simplified Linguistic Information Structure Annotation (LISA) guidelines described by Götze et al. (2007), where the assessment of ISR was considered at the level of the referential expression (i.e., the entire NP or PP) and not on the level of individual words (however, see Riester & Baumann, 2017 for a system that takes both levels into account).

The procedure used to annotate ISR was as follows. First, referential expressions were identified from the orthographic transcription in ELAN (without video or other annotations visible) and assigned a unique ID number, and then the referential expression was assessed for its information status, which was annotated on a separate tier. Discourse referents were considered new if they were not mentioned previously in the TED Talk nor could they be inferred through context or world knowledge. Referents were considered accessible when they could be inferred from context or could be assumed to be in the world knowledge of the listeners. An accessible referent could be inferred due to a previous referent via a number of relationships, such as a part-whole relation (e.g. *a building* / *its entrance*), a set relation (e.g., *the flowers in the garden* / *the flowers near the gate*), or entity/attribute

relations (e.g., *the flowers* / *their scent*)[29]. Finally, given referents were thus those that were explicitly mentioned previously in the TED Talk.

In order to account for complex embedded NPs and PPs, the annotation of ISR was done on two levels (that is, using a recursive approach, see, e.g., Riester et al., 2010; Riester & Baumann, 2013). For example, the phrase "the rights of content owners" would be considered one referent, or it could be parsed on the smallest level as two referents, *rights*, and *content owners*. Only this smaller level was considered for this current study. Additional aspects of information structure were annotated as well (namely contrastiveness, topic, and focus), but these go beyond the scope of the current study and thus will only be briefly mentioned in **section 5.4.** Reliability analysis of ISR annotations is pending.

5.2.6. Assessing multimodal cues to ISR

Not all multimodal speech acts are used to mark ISR. Our study adopted a common strategy to assess whether a prosodic cue can be considered as a marker of ISR by the presence or absence of a pitch accented syllable within the temporal bounds of the referential expression (e.g., "... and [she] **said** [it] would con**tin**ue on [a**cross** the **land**scape] **look**ing…", brackets indicate referential expressions and **bold** indicates pitch accented syllables). In other words, pitch accents that occur outside of the target referential expression (e.g., "said," "continue," "looking") were not considered to be markers of

[29] Examples taken directly from Götze et al. (2007)

ISR. However, this is considerably different from gestures, as previous research has shown that referential gestures may precede or follow their corresponding semantic information in speech (Graziano et al., 2020). Thus, a more holistic approach to assessing the gestural marking of ISR was adopted, largely following Desbreslioska et al. (2013), where each gesture was visually assessed to determine whether it impressionistically functioned to mark a discourse referent.

This assessment of whether a gesture was denoting a referent in speech was largely based on non-strict temporal association (e.g. Rohrer et al., 2019) and/or the semantic meaning conveyed by the gesture. Such an approach has clear advantages over automatically extracting simple temporal overlap, particularly in cases where strokes overlap the constituents of two referents, or (particularly for referential gestures) they precede or follow their corresponding referent. For example, **Figure 5.2** shows the pragmatic annotation of a non-referential gesture whose stroke overlaps two referents. The gesture co-occurs with the utterance *"*... people would think it was a hoax…". The stroke (indicated by the "Semantic_id tier") begins during the word "think" (upper left image) and spans the given referent "it" and ends during the second, new referent, "a hoax" (upper right image). An automatic extraction of temporal overlap would thus return that this single gesture stroke is "marking" two different referents with two different information statuses. However, a visual assessment gives the clear impression that the gesture is associating with, and consequently "marking", the second referent (indicated here as "SJ_302", a hoax). Thus, each

gesture stroke was visually assessed and the pragmatic domain tierset (see **Chapter 2** of this thesis) was annotated as having the pragmatic function of marking ISR. An additional "ref2_id" tier was created to identify which referent the gesture was perceived to be marking.

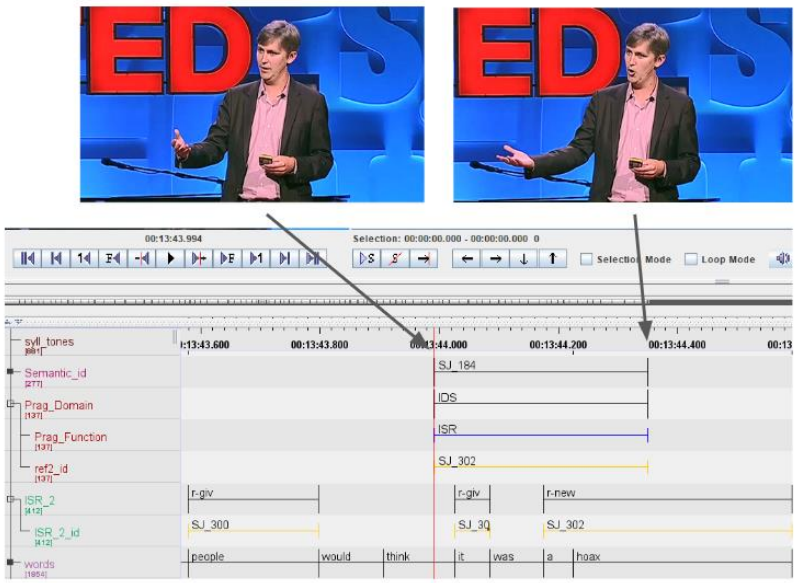

Figure 5.2: ELAN screenshot of the pragmatic domain tier set, and the annotation of ISR, taken from the English M3D-TED corpus by speaker SJ (Johnson, 2010) at 13:43.

5.2.7. Statistical analyses

In order to respond to the research questions, a series of Generalized Linear Mixed-effects Models (GLMM) with a poisson regression were run in R (R Core Team, 2022) with the *lme4* package (Bates et al., 2015). For each model, the *buildmer* package (Voeten, 2022) was used in order to determine the random effects structure that

returns the best fitting model. Mixed models that failed to converge or that overfit the data were rerun as simple Generalized Linear Models (GLMs). Significant effects were then assessed via omnibus test results, with a series of Bonferroni pairwise tests carried out with the *emmeans* package (Lenth, 2021). The following paragraphs explain each model in detail.

To respond to the first objective of the study to assess how multimodal cues (i.e., the combination of prosodic and gestural cues) are jointly used in our audiovisual data to mark ISR, a GLM with a poisson regression was run. The number of referential expressions was used as the dependent variable, and included a fixed factor of Information Status (3 levels: new, accessible, and given), of Multimodal Cue (4 levels: No Mark, Gesture Only, Pitch Accent only, and Both Gesture and Pitch Accent), as well as their two-way interaction. The model was offset by the total number of referential expressions for each category, so as to account for between-category differences[30].

For the second objective of the study, two models were run. In order to assess whether pitch accent type is involved in the marking of ISR, a GLM with a poisson regression was run with the number of pitch accented referential expressions as the dependent variable, with a fixed factor of Information Status (3 levels: new, accessible, and given), of Pitch Accent Type (7 levels: no accent, L*, !H*, H+!H*, H*, L*+H, L+H*), as well as their two-way interaction.

[30] glm(data = df, N_ref ~ InformationStatus*MultimodalCue, offset =log(total_refs), family = "poisson")

The model was offset by the total number of referential expressions for each category, so as to account for between-category differences[31]. In order to assess the effects of relative prominence, the syllable with the highest prominence score for each referent was identified (henceforth referred to as "max prominence score"). A linear regression was then run with the max prominence score for each referent as a function of Information Status and Multimodal Cue and their two-way interaction[32].

For the third objective of the study, two models were run. In order to assess whether gesture referentiality is involved in the marking of ISR, a GLMM was run with the number of gestures that mark ISR as the dependent variable, with a fixed factor of Information Status (3 levels: new, accessible, and given), of Gesture Referentiality (2 levels: Referential, Non-referential), as well as their two-way interaction. The model was offset by the total number of referential expressions for each category, so as to account for between-category differences, and the random effects structure included random intercepts for Gesture Referentiality by Speaker[33]. Finally, to assess whether gestures are sensitive to ISR in prenuclear accented positions, a GLMM with a poisson regression was run with the number of referential expressions marked with a prenuclear accent as the dependent variable, with a fixed factor of Information Status (3 levels: new, accessible, and given), of Multimodal Cue (2

[31] glm(data = df, N_ref ~ InformationStatus*PAtype, offset = log(total_refs), family = "poisson")
[32] lm(data = df, MaxProm ~ InformationStatus*MultimodalCue)
[33] glmer(data = df, N_gestures ~ InformationStatus*GestureReferentialty + (1 + GestureReferentialty | Speaker), offset = log(total_refs), family = "poisson")

levels: Pitch Accent Only, Both Gesture and Pitch Accent), as well as their two-way interaction. The model was offset by the total number of referential expressions for each category, so as to account for between-category differences, and the random effects structure included random slopes for Speaker[34].

5.3. Results

The annotation of the corpus resulted in a total of 1360 referential expressions, of which 228 referred to new referents, 362 referred to accessible referents, and 770 referred to given referents. Of the 1139 gestures that co-occurred with fluent speech (739 non-referential; 404 referential), 611 were considered to mark ISR (376 non-referential; 235 referential). In terms of pitch accentuation, of the 2033 pitch accented syllables annotated, 1006 were found to be produced within the constituent of a referential expression.

5.3.1. Gesture and prosody as joint cues for ISR

The goal of the current study was to better understand the joint multimodal marking of ISR. We hypothesized that pitch accentuation and manual gesture will often act together to mark ISR, specifically to mark newer referents, while given referents will most not receive multimodal marking. **Figure 5.3** shows the proportion of the different multimodal cues used for each ISR type (e.g., New, Accessible, Given). The overall results show that indeed

[34] glmer(data=df, N_prenuclear_refs ~ InformationStatus*MultimodalCue + (1 | Speaker), offset=log(total_refs), family = "poisson")

pitch accentuation and gesture are used together to mark newer information, while given referents are mostly not marked multimodally.

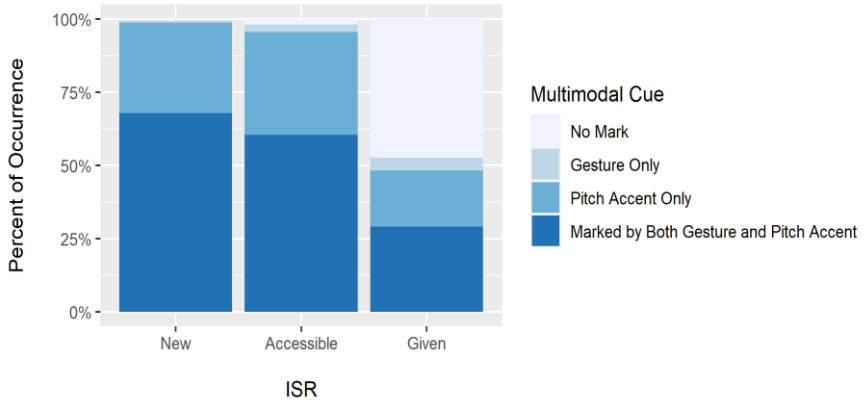

Figure 5.3: The observed proportions of multimodal cues by ISR.

The results of the GLM showed a significant main effect of Multimodal Cue ($\chi^2(3) = 562.17$, $p < .001$), as well as a significant interaction between Information Status and Multimodal Cue ($\chi^2(6) = 509.11$, $p < .001$). Post-hoc analyses of the interaction showed that new and accessible referents were significantly marked more often with multimodal cues than than given referents, which were more likely to not be marked multimodally (see **Table 5.1** for post-hoc pairwise comparisons).

Comparison	Multimodal Cue			
	Marked by Both Gesture and Pitch Accent	**Pitch Accent Only**	**Gesture Only**	**No Mark**
new-given	z = 8.649, p < .001	z = 3.396, p = .021	z = -2.182, p = .873	z = -5.699, p < .001
accessible-given	z = 8.299, p < .001	z = 5.084, p < .001	z = -1.273, p = 1	z = -8.521, p < .001
new-accessible	z = 1.09, p = 1	z = 0.838, p = 1	z = -1.646, p = 1	z = -0.986, p = 1

Table 5.1: Post-hoc pairwise comparisons of ISR for each multimodal cue.

When looking within each ISR category, it was revealed that both New and Accessible referents were significantly marked more by both multimodal cues than by pitch accent alone ($z = 5.15$, $p < .001$ for new referents; $z = 4.47$, $p < .001$ for accessible referents). This relationship was not found for given referents, which when marked multimodally could equally be marked by pitch accent alone or by both pitch accent and gesture. The entire database showed few cases where gesture marked ISR without pitch accentuation (showing no significant values across ISR categories).

5.3.2. Prosodic cues: Disentangling pitch accent type and relative degree of prominence

The second objective was to assess the relationship between pitch accent type, relative prominence, and ISR. **Figure 5.4** shows the relationship between ISR and Pitch Accent Type expressed as a proportion by Information Status type. The GLM returned a significant main effect of Pitch Accent Type ($\chi^2(6) = 723.36$, $p < .001$), and an interaction between the Pitch Accent Type and Information Status ($\chi^2(12) = 477.12$, $p < .001$). Post-hoc comparisons showed that given referents were produced without a pitch accent (deaccented) significantly more than new ($z = -6.373$, $p < .001$) or accessible referents ($z = -9.712$, $p < .001$). Given referents also received significantly fewer L*, !H*, and H* pitch accents than new or accessible referents.

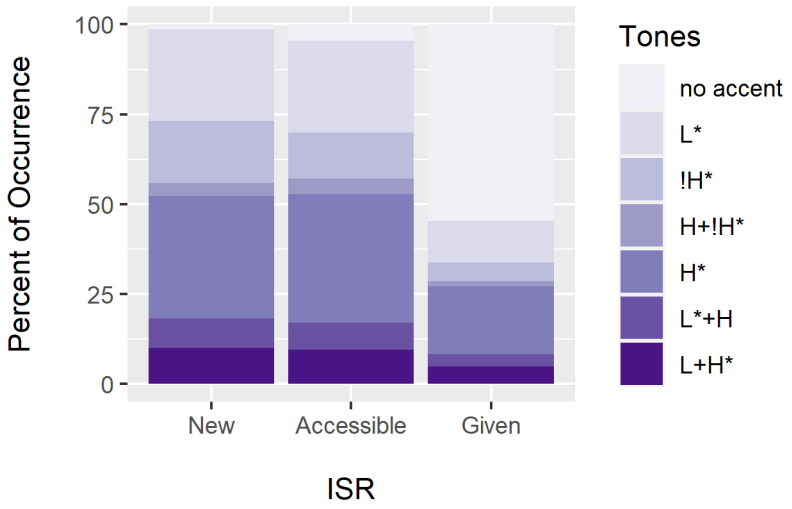

Figure 5.4: The observed proportion of pitch accent types by ISR

Regarding the role of relative prominence, the linear regression returned a significant main effect of Information Status ($F(2) = 31.323$, $p < .001$) and a significant main effect of Multimodal Cue ($F(3) = 660.254$, $p < .001$) yet no significant interaction between the two (see **Figure 5.5**). Pairwise comparison of the main effect of Information Status showed that both new and accessible referents were perceived to be significantly more prominent compared to given referents ($t(1166) = 3.435$, $p = .002$; $t(1166) = 5.185$, $p < .001$, respectively). Post-hoc comparisons of the main effect of Multimodal Cue showed that referents marked by both pitch accent and gesture were perceived to be significantly more prominent than those produced with a pitch accent alone, both of which were in turn significantly more prominent than referents parked either with gesture only or no multimodal marking (with no significant differences between the latter two, see **Table 5.2** for all potential pairwise comparisons for the main effect of Multimodal Cue).

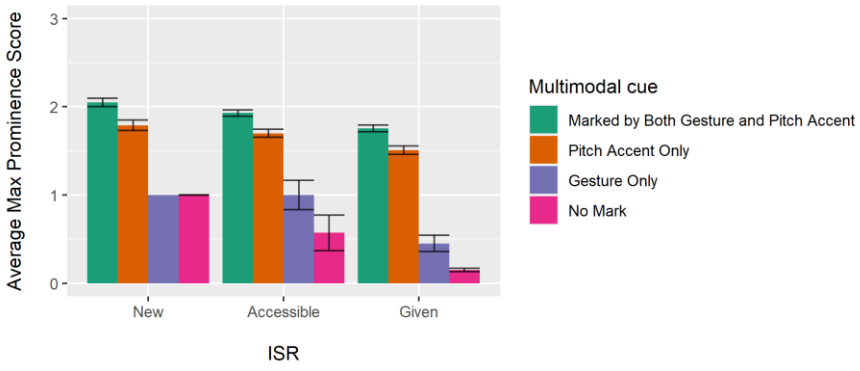

Figure 5.5: The average prominence of referents as a function of their information status and multimodal cue (error bars represent standard error).

	Marked by Both Gesture and Pitch Accent	**Pitch Accent Only**	**Gesture Only**	**No Mark**
Marked by Both Gesture and Pitch Accent	–	t(1166) = 6.869, p < .001	t(1166) = 6.648, p < .001	t(1166) = 10.737, p < .001
Pitch Accent Only		–	t(1166) = 5.128, p < .001	t(1166) = 8.687, p < .001
Gesture Only	–	–	–	t(1166) = 1.192, p = 1
No Mark	–	–	–	–

Table 5.2: Post-hoc pairwise comparisons for the significant main effect of Multimodal Cue as a predictor of maximum prominence score

5.3.3. Gestural cues: Differences by gesture type and their role in prenuclear positions

The third objective was to better understand the effects of gesture referentiality as a multimodal cue for ISR, and as a cue in prenuclear accented positions. Specifically in terms of gesture referentiality, the GLMM did not reveal any significant effects nor

an interaction between Gesture Referentiality and Information Status (see **Figure 5.6**). However, when assessing gesture production at prenuclear accented positions, the GLMM showed a significant main effect of Multimodal Cue ($\chi^2(1) = 8.182$, $p = .004$) and a significant interaction between the Multimodal Cue and Information Status ($\chi^2(2) = 6.804$, $p = .033$). The post-hoc analysis revealed that accessible referents which receive a prenuclear pitch accent were significantly more likely to also receive a gesture ($z = 2.83$, $p = .042$). This relationship was not found for new ($p = .074$) or given ($p = 1$) referents that were marked by prenuclear pitch accents (see **Figure 5.7**).

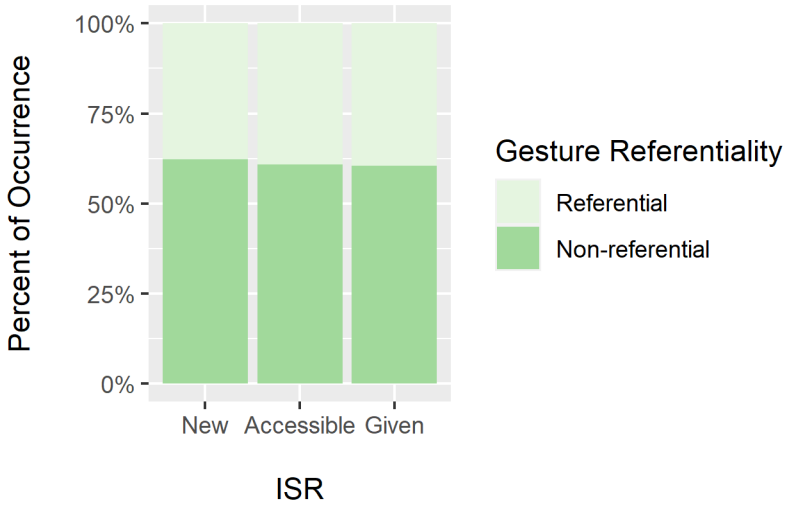

Figure 5.6: The observed proportion of gesture types by ISR.

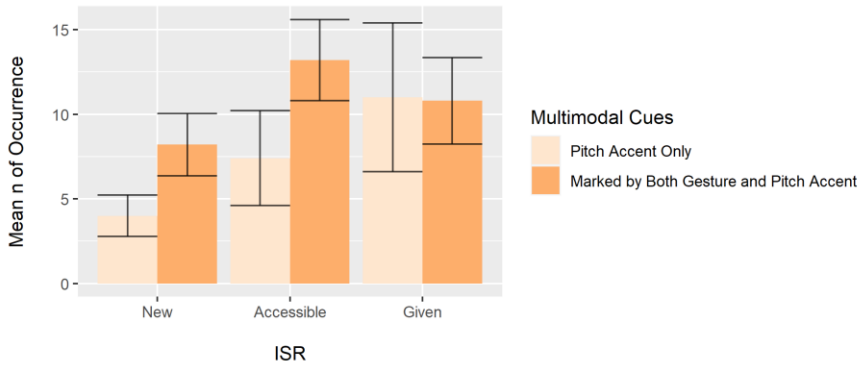

Figure 5.7: The average number of occurrences of referents with prenuclear pitch accents by speaker as a function of whether they co-occurred with gesture or not (error bars represent standard error).

5.4. Discussion and conclusions

The goals of the current study were to assess the joint role of gesture and prosody in marking ISR in discourse, to assess the mapping between information status and pitch accent type, prominence, gesture type, and to evaluate the gestural marking of ISR in prenuclear positions. We found that pitch accentuation and gesture jointly mark information structure, with new and accessible referents receiving mostly a double marking by both prosody and gesture, and with given referents receiving multimodal marking significantly less than the other referent types. Crucially, the results of the current study did not find any evidence of a one-to-one mapping between neither pitch accent type nor gesture type and ISR. Interestingly, relative prominence in general indeed seems to be in function of ISR, in the sense that new and accessible referents were marked with pitch accents that were perceived to be

significantly more prominent than given referents. Finally, when referents are marked by prenuclear pitch accents, only accessible referents are also more likely to co-occur also with a gesture. The following paragraphs will discuss each of our principal findings in order.

The fact that pitch accentuation and gesture jointly mark the ISR is not entirely surprising. First, the close temporal relationship between prosody and gesture has already been shown in the literature (e.g., **Chapters 1**, **3, and 4** of the current thesis; Shattuck Hufnagel & Ren, 2018 for a review, among many others). However, a closer look at these results reveals some interesting findings. Importantly, new and accessible referents were significantly more likely to associate with both a pitch accent plus a gesture (66.67% and 59.39%, respectively) than with pitch accent alone (32.02% and 36.19%, respectively), while given referents were most likely to receive no joint multimodal marking (with 49.87% receiving no multimodal marking, compared to 0.8% of new referents and 1.9% of accessible referents receiving no marking). Furthermore, pitch accentuation may be a more robust marker in general than gesture, as referents that are marked by only one mode tend to be marked by pitch accents only rather than gesture only. This is again in line with the common observation that a majority of gestures co-occur with pitch accents, but it is not necessarily true that the majority of pitch accents are accompanied by gesture.

These results are in line with aspects of McNeill's theories on Communicative Dynamism (1992) and Givón's "Principle of

Quantity" (1983), where less accessible information (in this case, new referents) should receive more marking material. An interesting observation that is worthy of mention is the fact that approximately half of given referents still receive some sort of multimodal marking (26.36% receiving both pitch accent and gesture, 19.74% receiving a pitch accent only, and 4% receiving gesture only), which could be considered a relatively high number, as the literature has documented that given referents are often deaccented (e.g., Pierrehumbert & Hirschberg, 1990; Ladd, 2008, among many others) and co-occur with few gestures, especially when produced with reduced/pronominal forms (e.g., McNeill, 1992; see also works by Debreslioska and colleagues). In our view, this might be due to the fact that academic-style speech tends to elicit very prosodically and gesturally emphasized speech with high density of pitch accentuation and gesture. Moreover, a number of studies have found similar results where given referents indeed receive a pitch accent or co-occur with a gesture (Mücke & Grice, 2014; Im et al., 2018). Some explanations could be that the given referent is interacting with other levels of information structure, such as occurring within focal positions (e.g., Féry & Kügler, 2008), they may be involved in marking contrast, (or alternatively, to show that modifiers are not contributing to an alternative reading, Riester & Piontek, 2015) or they may be reintroduced rather than maintained referents (Debreslioska & Gullberg, in press). Apart from information-structural relations, other linguistic aspects may be at play, such as rhythmic constraints (e.g., Baumann et al., 2007; Calhoun, 2010a) or the fullness of the referential expression (e.g.,

Debreslioska & Gullberg, 2013). One example of such complex interactions can be demonstrated in an extract taken from the English M3D-TED database, where speaker EG says at 06:53, "which is great, because the Romans did not actually think that a genius was a particularly clever individual. They believed that a genius was this sort of magical, divine entity" (Gilbert, 2009). In both cases, the word *genius* is a given referent. However, the speaker produces a non-referential gesture each time she utters the word *genius*. Interestingly, she produces the gesture in two distinct locations, initially on her left, and at the second occurrence, on her right, which could be seen as encoding a contrastive reading (e.g., Gullberg, 1998, p. 148).

The second set of findings showed how traditional claims that have related categorical measures of pitch accent type and/or gesture types to ISR marking do not hold for our data. In terms of pitch accents, recent studies have adopted a "probabilistic" approach, where newer information is merely more likely to be marked with greater prominence through pitch accentuation (e.g., Mücke & Grice, 2014, see also Cangemi & Grice, 2016). However, the results from the current study did not show a probabilistic relationship between pitch accent type and ISR category. This is partially in line with recent work (Baumann et al., 2021, regarding prenuclear accentuation), which has highlighted how relating prosodic prominence, pitch accent type, and informativeness is much more complex than previously assumed. The current study took a paradigmatic approach in the analysis of pitch accent type, looking at referents and their pitch accentuation across the entire database.

However, by focusing on local relationships within, for example, the intonational or intermediate phrase (IP or ip, respectively) may elucidate clearer relationships between the encoding of pitch accent type and ISR. Future studies should not only control for other aspects of speech as those mentioned above regarding the accentuation of given referents (e.g., rhythmic constraints, morphosyntactic effects) but also consider a syntagmatic approach, looking at the mapping of pitch accent type to ISR at a local level.

Importantly, the relative prominence of syllables was indeed annotated on the level of the IP by using exclusively audio information (see **subsection 5.2.4.2**). The results show how the newer the referent, the more prominent it tended to be. This is particularly interesting considering when looking at the relative prominence of pitch accented syllables by pitch accent type, the trend seems to follow what has been described in the literature (namely, syllables with a L* pitch accent tended to receive lower prominence ratings, followed by !H*, H*, and finally rising pitch accents with the highest prominence ratings, for similar results at least from those at the upper end of the scale, see Cole et al., 2018). The apparent mismatch between relative prominence, pitch accent type, and ISR further suggest that a paradigmatic approach may be useful in elucidating this relationship, where potentially a L* pitch accent may be produced with greater relative prominence within the phrase to achieve the pragmatic goal of marking ISR. Interestingly, this analysis also showed how syllables that were produced with gestures were perceived as being more prominent than those produced without gesture. It is important to bear in mind that these

annotations were done without access to the video, and thus the annotator was unaware when the speaker was producing gesture (for more about how the production of gesture may impact the production and perception of prosodic prominence, see, e.g., Krahmer & Swerts, 2007). Though no specific interaction was found, it seems that across the board, gesture is being used to reinforce prosody as a visual prominence marker for ISR, ultimately showing how gesture and speech prosody are integrated to achieve pragmatic goals in communication.

In terms of gesture referentiality, again no interaction was revealed indicating that non-referential gestures are no more particularly suited to be markers of new information than referential gestures. This finding is particularly important as it clarifies often contradictory claims in the literature. Some studies have assigned non-referential gestures a particular function of focus marking (e.g., Loehr, 2004, 2012), introducing new topics or characters (McNeill, 1992), among others. Alternatively, McNeill's theories of CD have posited that deictics and non-referential gestures should occur when utterances contain low amounts of CD, and more semantically rich iconic and metaphoric gestures should occur when CD is high. However, the current findings did not find any differences between gesture types for the marking of the ISR. Of course, ISR is merely one aspect of information structure, and perhaps by adopting more inclusive aspects such as informativeness (which integrates contrastiveness on the same level, Baumann et al., 2022) or including interactions with other levels of information structure (such as referents within focus or rhematic constituents) may offer a

more fine-grained measure of CD. Thus, future studies should take more levels of IS into account. Other aspects of morphosyntax may also play a role. For example, one study by Ferré (2010) investigated the gestural marking of focus through marked expressions in syntax (i.e., predicate, argument, and sentence focus, as per Lambrecht, 1994) and prosodic focus in French. The author found that in French, non-referential gestures tended to associate with focus marking when it was also marked prosodically. However, when focus was marked syntactically, metaphoric gestures were often used. Thus, future studies may take account of language specific aspects of various prosodic and morphosyntactic strategies to the marking of information structure.

Regarding the final objective, it seems that when referents are marked with a prenuclear pitch accent, it is only accessible referents that are more likely to co-occur with a gesture as well. While previous studies on the marking of ISR in prenuclear positions have noted that acoustic parameters of those pitch accents may differ in terms of information status (e.g., Kügler & Féry, 2008; Braun, 2006; see also Baumann, 2021), the results of the current study suggest that gesture production continues to be sensitive to ISR. Thus, in cases where pitch accentuation may function principally for the rhythmic organization of speech (as per Calhoun, 2010b), gesture may step in to take on the role of marking ISR. Furthermore, a recent study by Debreslioska & Gullberg (2020b) have found gesture to associate more with accessible referents over new referents. The authors suggest that this may be due to the linguistic encoding of accessible referents with definite nominals

(much like given referents). Speakers may use gesture as a strategy to signal to the addressee that it is in fact not a given referent, but something that is new to the discourse (though inferable in some way) and thus it should be treated as a new referent. Though the current study did not find that gestures associated more with accessible referents than new ones across the board, it did indeed find this to be the case when accessible referents were in prenuclear positions. The current results can be thus interpreted as adding prosodic evidence in support of the hypothesis by Debreslioka & Gullberg (2020b), namely that when a referent is in prenuclear position (a position often associated with given referents), the gesture may be associating with it to indicate to speakers that it is indeed to be treated as a newer referent.

Thus, the view that emerges from these results allows us to better understand the interconnections between prosody and gesture in the multimodal marking of ISR. The results do not support the view of a one-to-one mapping between multimodal cues, in the sense that gesture association is directly dependent on prosodic prominence. Rather, they bring a more nuanced view of the phonological and pragmatic synchrony rules, showing that direct ISR-Gesture relations emerge and gesture is not parasitic on prosodic structure. More work will be needed to disentangle the complex relationship between gesture production, prosodic structure, and their contribution to pragmatic meaning.

The current study has some limitations that should be addressed in future studies. First, the study only investigated the ISR using a

simplified three-level distinction between new, accessible, and given referents (using the LISA system as described in Götze et al., 2007). It may be interesting for future studies to use more precise coding systems which account for more precise levels of information status. For example, the LISA coding system indeed proposes a more detailed description of different subtypes for more complex labeling. Similarly, the RefLex system (Baumann & Riester, 2017) accounts for multiple types of accessible referents, separating out those that can be inferred from situational context from those inferred from world knowledge. RefLex also controls for aspects such as co-referentiality (the use of, e.g., synonyms that refer to the same discourse referent — here coded as accessible — could be distinguished on a referential and lexical level). Future studies may also wish to account for interactions between levels of information structure (e.g., given referents in focus position, contrastiveness, etc.). Second, future studies may wish to assess these relationships in other genres of discourse. While TED Talks reflect a natural, semi-spontaneous style of speech, it is produced in a very specific style and context. For example, academic-style speech tends to elicit more non-referential gestures than referential ones (e.g., Shattuck-Hufnagel & Ren, 2018), and this may be different in tasks which are less narrative in nature (e.g., a map-direction task, a picture description task, etc.). More studies involving natural multimodal interactions and spontaneous conversational speech would be welcome too (e.g., Holler et al., in press). Additionally, the prosodic annotation was carried out by a single annotator, and thus could benefit from having multiple

annotators (e.g., through Rapid Prosodic Transcription, or RPT, Cole & Shattuck Hufnagel, 2016).

All in all, the present study has shown that pitch accentuation and manual gesture work together to mark the ISR in academic style speech. While specific pitch accent types or gesture types were not found to be particularly suited for the marking of ISR, the study did reveal that relative prominence still appears to be marking ISR, and that gesture continues to act as a multimodal cue to information structure even in prenuclear accented positions. Importantly, these results show that gesture is not merely parasitic on pitch accentuation, but exhibit a complex relationship mediated by pragmatic meaning. Thus, gestures appear to have a strong role as visual markers of prominence which is integrated not only temporally, but coordinated pragmatically with speech prosody to mark ISR. This highlights the complex relationship gestures have with speech prosody, and how multimodal communication should be assessed independently across dimensions of morphosyntax, prosody, and gesture.

6

CHAPTER 6: GENERAL DISCUSSION AND CONCLUSIONS

6.1. Summary of findings

The current thesis had two main goals. First, to propose a novel multidimensional approach to the study of gesture and to better understand the prosodic and pragmatic characteristics of both referential and non-referential gestures in two typologically different languages, namely French and English. The second objective which was investigated across three empirical studies represents a refinement of McNeill's (1992) phonological and pragmatic synchrony rules. The thesis contained four studies, where the first study laid out the proposal, and was followed by three empirical studies that applied the M3D proposal in order to achieve the second objective of the study.

The study in **Chapter 2** described the MultiModal MultiDimensional (M3D) labeling system for audiovisual corpora. Through a review of 10 currently available multimodal labeling systems, we found that only half of the systems apply concepts related to gestural phasing that are widely accepted in the field. Moreover, there is a wide variety of approaches to assessing the semantic meaning of a gesture and also to classifying gesture types. Importantly, very few systems account for the multiple pragmatic meanings a gesture may convey, and very few systems describe any sort of prosodic coding. While access to the descriptions of the labeling systems is relatively open, the descriptions themselves are oftentimes not very explicit and thus their applicability to novel corpora by labelers inexperienced with the coding system remains problematic.

The M3D system addresses the aforementioned issues. First, the M3D approach takes on a more comprehensive perspective of multimodal language, by including many different aspects of language such as speech, prosody, manual gesture, head movements, facial expressions, gaze, etc. Crucially, it is one of the few multimodal labeling systems that aims to better understand the different modes, or meaning-making strategies that speakers use when engaging in communication by coding separately the speech (i.e., the morphosyntactic utterance), the prosodic, and the gestural channels. Second, it proposes a tripartite dimensional approach to the annotation of gesture, independently accounting for gesture form, the prosodic dimension of gesture, and the meaning dimension of gesture (i.e., their semantic and pragmatic contributions). Finally, it openly offers multiple resources (e.g., a labeling manual, training resources, an annotated corpus) and guidance on how to thoroughly assess gestures across their dimensions. Crucially, the reliability for key aspects of the coding system were shown to be quite high, even when coding aspects as non-mutually exclusive dimensions. It is argued that this approach refines and integrates two leading conceptual theories in the field (i.e., McNeill, 2005; Kendon, 2004; 2017) by incorporating the referentiality divide by McNeill in the M3D semantic dimension tiers and the pragmatic functions of gestures largely described by Kendon in the M3D pragmatic dimension tiers. By having an integrative theoretical approach to multimodal and gesture labeling and by offering standard annotation practices, M3D has the

potential to foster interdisciplinary work and allow for easier comparison between studies in the field.

Importantly, the second goal of the thesis was to assess the pragmatic and prosodic characteristics of gestures, which refine the phonological synchrony rule. By adopting M3D and applying it to two typologically different languages in terms of prosody (i.e., in their prominence-marking strategies, where French is a demarcative-based language and English, a prominence-lending one), we are better positioned to assess whether gestures are directly dependent on pitch accentuation and prosodic structure. That is, whether they associate with pitch accents equally, regardless of their position in phrasal prosodic structure.

The study in **Chapter 3** investigated the association of referential and non-referential gestures with pitch accentuation, as well as rhythmic relationships between the two modes in a corpus of five French TED Talks (containing over 37 minutes of multimodal language annotations, including over 1500 gestures). Specifically, it has assessed for the first time gesture-speech synchrony on the level of the pitch accent in French, accounting for gesture type (referential vs. non-referential) and phrasal position (phrase-initial IA vs. phrase-final FA). The results showed that strokes largely associated with pitch accented syllables, while apexes were much more variable in terms of association patterns (based on the alignment criteria established in the current studies). Furthermore, no appreciable differences were found between the association rates of referential and non-referential gestures. Crucially, the results of

the analysis found that when two potential pitch accents were present within the prosodic phrase, gestures preferably associated with the IA, at the left edge of the prosodic phrase.

Moreover, studies on the rhythmic production of subsequent gestures have suggested that their temporal production is largely independent of pitch accentuation. However, in French, accentuation is closely related to phrasing, where particularly the FA marks the last full syllable in the AP. Importantly, studies have shown that the duration of the AP, regularly indicated by FA, is key in the perception of speech rhythm in French (e.g., Pasdeloup, 1992; Mertens, 1992; Delais-Roussarie, 1995; Astésano, 2001). Thus, the study also investigated the rhythmic relationship between speech and gesture production, by assessing rhythmic groups of subsequent gestures (RGGs) and their association with speech rhythm, accounting for prosodic phrasing for the first time. The results demonstrated that when assessing RGGs, they tended to be more often non-referential in nature than referential, with no differences in isochrony being observed between the two types of RGGs. Additionally, no direct correspondence between RGGs and pitch accentuation patterns was found, showing that RGGs are not closely related to the instantiation of pitch accents. However, duration of the prosodic phrase indeed predicted the interval between subsequent gestures, showing that as AP duration increased, as did the interval between subsequent RGG apexes. These results show for the first time that even though RGGs are loosely bound by complex prosodic structure, they are independent of pitch accentuation prominence and should be studied separately. Future

studies might want to investigate the intricate relationship between RGGs and prosodic structure and RGGs as promoting an independent rhythmic dimension.

The study in **Chapter 4** investigated the temporal association of referential and non-referential gestures with pitch accentuation in a corpus of five English TED Talks (representing over 23 minutes of multimodal language annotations, including over 1150 gestures). Crucially, it is the first time that the relationship has been assessed while taking into account effects of complex prosodic structure, in terms of pitch accent nuclearity, phrasal position, and relative degree of prominence. Similar to the study in **Chapter 3**, strokes were found to reliably align with pitch accented syllables while apexes were more variable in their alignment patterns, with no differences being observed between gesture types. Importantly, when gestures occurred within an intermediate phrase which contained prenuclear accents, gestures associated more with prenuclear accents than with the nuclear ones. Moreover, this was not driven by an effect of relative prominence, suggesting that there is a structural effect where gestures tend to occur early in the ip, most frequently marking the left edge. Crucially, this finding coincides with the IA effect found in French, showing that in both languages, gestures have a tendency to mark the left edge of a prosodic phrase.

In sum, the results from **Chapters 3 and 4** have shown that phrasal prosodic structure influences gesture-speech phonological synchrony, particularly by finding an edge-initial strengthening

effect whereby the left edge of the prosodic phrase constitutes a pole of attraction for gestures. These results crucially help refine the phonological synchrony rule. Importantly, the pragmatic synchrony rule also needed some refinement, as little was known about how gesture and prosody jointly function to mark the information status of referents (ISR).

Therefore, the study in **Chapter 5** investigated the multimodal marking of the ISR in English TED Talks, cosidering both modes of communication (i.e., gesture and prosodic prominence) for the first time, as well as taking into account phrasal prosodic structure. The main findings from this study showed that new and accessible referents were marked by both pitch accent and gesture more often than a single cue alone, while given referents were significantly more likely to not be marked by either cue. This finding suggests that both modes of communication (pitch accentuation and gesture) jointly function to mark the information status of referents (ISR). Specifically in terms of prosodic marking, no clear relationship between ISR and pitch accent type was found, while relative prominence seemed to be a more robust marker of ISR. Finally, in terms of gesture, no significant differences between gesture types were found. Crucially, in prenuclear pitch accented positions, an interaction between gesture and prosody was found for the first time, showing that gestures marked accessible referents significantly more than given or new ones, playing a complementary role with pitch accentuation.

In the upcoming sections, I will discuss these findings in relation to the previous literature and show how they contribute to the field by refining the phonological and pragmatic synchrony rules by McNeill (1992) and our understanding of how these speech prosody and gesture interact in multimodal language, as well as comment directions for future work.

6.2. The value of M3D

The M3D proposal fulfills a current need within research in multimodal communication for a more standardized approach to multimodal data annotation that (a) incorporates recent advances in the gesture field regarding the multidimensional analysis of gestures; and (b) allows for annotations to be interpretable across investigation sites, and flexible enough to meet individual researchers' needs. Specifically, it systematically covers the three dimensions of gesture, namely the form dimension, the prosodic dimension, and meaning dimension. Importantly, M3D disentangles gesture form from semantic and pragmatic meanings, as well as prosodic characteristics, proposing a tripartite dimensional analysis that is to be performed for all gestures. Importantly, such an approach bridges the gap between two of the most prominent theoretical approaches in the field of gesture studies (i.e., Kendon, 2004; McNeill, 1992) through the incorporation of two important novelties: a) the assessment of semantic meaning in terms of gesture referentiality and the possibility of coding referential gestures in terms of potentially-overlapping dimensions rather than mutually-exclusive categories (in accordance with McNeill's (2006)

proposal) and b) the possibility of annotating and thus exploring a range of non-mutually exclusive pragmatic meanings in a reliable manner. Through measures of inter-annotator agreement, we have shown that these key parts of the M3D system can be reliably coded.

The findings from the three empirical studies in **Chapters 3**, **4**, **and 5** largely lend support to some central claims in the M3D approach to gesture labeling. First, the results from these three chapters show how language is multimodal in nature, where speakers make use of multiple meaning-making strategies for communication, and importantly, that these modes interact at various levels, both in terms of their temporal co-production, as well as in marking meaning (in terms of ISR). M3D is among the few multimodal annotation systems that help elucidate such relationships in multimodal communication. The results of this thesis highlights how language is indeed multimodal, and researchers in various subfields of linguistics would benefit from more interdisciplinary approach that work to elucidate the complex interactions between various aspects of language, including not only speech prosody and gesture (as was the focus in the current thesis), but also in the fields of pragmatics, syntax, and semantics. M3D thus offers a framework within which all of these aspects of language can be studied, ultimately fostering interdisciplinary approaches and collaborations across subfields of linguistics. By adopting such an approach as proposed by M3D, researchers have an appropriate tool for a more comprehensive and multimodal assessment of human language and how it functions in communication.

Second, the findings in **Chapters 3**, **4**, **and 5** also lend support to the tripartite dimensional assessment of gestures, namely their form, prosodic, and meaning dimensions. By independently assessing these three gestural dimensions, the results of the present thesis can refine the traditional characterization of "beat" gestures as special prosodic markers of prominence and discourse-pragmatic functions that are different from gestures that are referential in nature. The results of the three studies show that in both M3D-TED corpora gestures of all types, be they referential or non-referential, generally associate with pitch accentuation both in French and in English. This finding reinforces the evidence reported in earlier studies (i.e., Pouw & Dixon, 2019b; Shattuck-Hufnagel & Ren, 2018) and moves the phonological synchrony rule back to the center as being applicable for all gestures regardless of their referentiality. Thus, the approach espoused by M3D that assesses the different dimensions of gesture independently helps demythify the unfounded idea that non-referential gestures have a special relationship with prominence in speech.

Furthermore, it is important to mention that non-referential beat gestures were also initially characterized as beating musical time (i.e., to be produced in a rhythmic fashion). Interestingly, while we found both referential and non-referential gestures being rhythmically produced, the results indeed suggest that non-referential gestures are more likely to be produced in a subsequent, rhythmic fashion. However, it is important to note that the assessment of beat-like groups of gesture was done perceptually, that is, if a group of gestures were perceived as forming a rhythmic

group, they were considered as such. It is possible that the lack of semantic content manifesting in gestural form may have led to a bias towards perceiving non-referential gestures as being more rhythmic than referential ones. Future studies on the rhythmic production of subsequent gestures may thus consider taking more objective measures, such as relative time differences between subsequent gestural phenomena to identify "groups of gestures" in a more quantitative way. In any case, these findings lend support to the idea that gestures can have prosodic characteristics - they can have a beat-like, "prominence-lending" function that can group together to form groups of subsequent gestures which mark rhythm, and once again this is not dependent on gesture referentiality.

Finally, the M3D view that gestures are not *either* semantic *or* pragmatic in their meaningful contribution to speech has been backed up by the results obtained in **Chapter 5**. As we have seen, both referential and non-referential gestures signal ISR at similar rates. More recent views have indeed described how gestures can be multifunctional in that they can express multiple semantic or pragmatic meanings in a holistic fashion. McNeill (1992) acknowledged how a referential gesture conveys multiple aspects of its referent (e.g., showing shape, size, and patch of movement holistically). Similarly, a single gesture may fulfill multiple pragmatic functions simultaneously (e.g., Lopez-Ozieblo, 2020). M3D thus integrates McNeill's views on referentiality and Kendon's pragmatic approach to be more in line with these more recent views on the multilevel view of conveying meaning In general, these findings reinforce the central idea that all gestures

should be characterized in terms of three largely independent dimensions: their form, their prosodic characteristics, and their semantic and/or pragmatic contribution to speech.

The paragraphs above have laid out how the M3D approach represents a valid approach that has the advantage (a) to assess gesture properties across three largely independent dimensions, and (b) to explicitly recognize that gestures represent only one mode of communication, which may interact with other modes in language. M3D makes another important contribution to the field in offering a multimodal labeling system that is highly accessible, explicit, and applicable to a wide variety of corpora. Importantly, it is the only labeling system to our knowledge that offers explicit, step-by-step guidelines in how to annotate multimodal corpora, with examples and suggestions for ambiguous cases, tips for workflow, and a template for labelers to begin applying M3D immediately. All of these resources offer a key opportunity to the field of gesture research, in that it has the potential to bring gesture researchers together regardless of their particular theoretical approaches, enabling collaborations across the field and ultimately producing comparable and reproducible research.

6.3. Refining the phonological synchrony rule

The current section will try to assess the findings in the present thesis that are helpful to refine McNeill's phonological synchrony rule. The phonological synchrony rule stated that gesture strokes occur just before or coincide with the "phonological peak syllable

of speech" (McNeill, 1992, p. 26). However, most studies had focused only on prominence without taking into account phrasal structure. **Chapters 3 and 4** of the thesis contributed two empirical studies on two typologically different languages from the point of view of prosodic structure that further our understanding of the prosodic association of gestures. In terms of gesture-speech synchrony, most early observations have described nuclear pitch accents as being a key gesture anchoring point (Kendon, 1980). Two main areas were assessed, namely the temporal association between gesture and pitch accentuation, but also the role of phrasal prosodic structure. While it is often attested that most gestures tend to be temporally associated with pitch accents, not all pitch accents receive a gesture. Shattuck-Hufnagel & Ren (2018) comment that this is indeed a larger question that needs further study: why are some gestures produced without prominence in speech, and why are some pitch accents produced without co-accompanying gesture? **Chapters 3 and 4** in this thesis are set to answer some of these questions.

6.3.1. The association of strokes with pitch accented syllables: High levels of synchronization across gesture types

The studies in **Chapters 3 and 4** of this thesis have gone on to reconfirm previous studies showing that individual gesture strokes and pitch accentuation align at very high rates for both French and English (e.g., Karpiński et al., 2009; Nobe, 1996; Shattuck-

Hufnagel & Ren, 2018). Interestingly, this finding has been shown for the first time in French, where pitch accentuation is not prominence-lending. Importantly, for both languages, there was no difference between referential gesture types. This finding shows how there is not a single gesture type that specifically associates with speech prominence more than other gesture types (as per the traditional conceptualization of "beat" gestures). These empirical findings lend further evidence to the initial phonological synchrony rule in that all gestures regardless of their semantic nature typically associate with prominence in speech, and validate the approach proposed in M3D to which prosodic properties of gesture, and their association with speech prominence should be assessed independently of their referentiality properties.

Moreover, the findings on gesture-speech temporal integration add to our understanding of gesture-speech synchrony in typologically different languages. As previously mentioned, most studies on gesture-speech temporal association have focused on languages where pitch accentuation is prominence-lending, such as English, Italian, Dutch, German, and Catalan, etc. One study by Fung & Mok (2018) on pointing in Cantonese found that *f0* did not play a role in gesture-speech synchronization. Instead, they found that gesture apexes regularly occurred on the first syllable of the word carrying prosodic stress (encoded via syllabic lengthening) regardless of whether the first syllable was prosodically stressed or not. The results of the study in **Chapter 3** indicate that even in French, where pitch accentuation is said to be more demarcative, they continue to act as a prosodic anchor for gesture production.

This may be because even though pitch accentuation is not necessarily prominence-*lending*, they still convey a degree of prominence in the speech stream that helps build up rhythmic patterns in the language. The results force us to reconsider whether French can be characterized as being vastly different from other (prominence-lending) languages in terms of prosodic structure. By the same hand, we have shown how strokes closely associate with accentual structure (at rates around 80% in both languages). Thus, gestures help visualize accentual structure in typologically different languages.

6.3.2. The role of the apex in the temporal relationship between gesture and pitch accentuation

In contrast with the findings on the high levels of temporal alignment between gesture strokes and pitch accentuation in both languages, the databases in the current thesis could not show the same tight relationship with apexes. Indeed, the finding that apexes are tightly correlated with pitch peaks has largely been found in experimental settings where participants are explicitly asked to produce a beat or a pointing gesture on target words. Results from studies investigating natural speech have varied. Loehr (2004; 2012) found that apexes occurred on average within 17 ms of a pitch accent. However, the standard deviation is quite large (341 ms), suggesting that in fact, many apexes may fall outside of the bounds of the pitch accented syllable as the standard deviation is larger than the average word length in his data (reported to have a mean duration of 227 ± 132 ms).

A recent study by Pouw & Dixon (2019b) used a narrative retelling task and recorded movements via motion tracking. They found that the peak velocity of the stroke occurred on average 39 ms (± 454) before pitch peak, and peak deceleration occurred on average 44 ms (± 424) after the pitch peak. The authors conclude that the two measures closely synchronize with pitch peaks as the confidence interval resulting from statistical tests contained 0. Regardless, given the high standard deviation, this suggests again that there are many apexes occurring outside of the interval of the pitch accented syllable. Furthermore, the study only used the acoustic measure of pitch peak and did not do any phonological assessment. That is, they did not annotate syllable boundaries nor assess whether these peaks correspond to actual pitch accents). Two studies on natural speech that did carry out such assessments were those by Yasinnik et al. (2004) and Esposito et al. (2007). The former indeed found alignment rates similar to those reported in the current study (i.e., ~65%) for monosyllabic words. Their results for polysyllabic words are unclear, as they report "Of the 130 hit- aligned words, 117 or 90% also contained a Pitch Accent" (p. 13), suggesting that an apex could occur on a non-pitch-accented syllable within a pitch accented word. The latter study very clearly assessed apex alignment within the boundaries of pitch accented syllables in two monologues by two native Italian speakers. While they report high rates of alignment across multiple articulators (including the hands, head, eye brows, and shoulder shrugs; 78% and 84% alignment for their two speakers), a closer look at their published data regarding manual co-speech production showed lower rates and a high degree

of inter-speaker variability, where one speaker produced many manual gestures (with apexes aligning at a rate of 63%), while the other speaker only produced three manual gestures, of which two aligned. Again, these results suggest lower alignment rates than in laboratory studies.

The discrepancy between the results from laboratory studies and studies on natural speech could be viewed in terms of a matter of precision. Specifically, the current database showed that most apexes that did not co-occur within the bounds of a pitch accented syllable were produced on an immediately adjacent syllable. Thus in laboratory studies where participants are asked to gesture on target words, they are more aware of their gestural productions and thus consciously control their productions in a more precise manner. Gesture production in natural speech is much more spontaneous and often occurs without speakers even realizing it. In such contexts, a more "loose" association may surface. This finding thus has important conceptual and methodological implications. First, it adds nuance to the pervading idea that gesture apexes are anchored in pitch accented syllables. While the two may be closely coordinated, this coordination is not as tight in spontaneous speech. Methodologies that account for such a loose association may be a better approach. For example, some researchers argue that synchronization may be better conceptualized as two phenomena co-occurring at a regular distance from each other, regardless of whether occur simultaneously or they fall within discrete boundaries (such as a pitch accented syllable) (Leonard & Cummins, 2011; Turk, 2020). Future studies regarding apex-pitch

accent association on natural speech data may take more gradient approaches, estimating time differences regardless of boundaries (much like what has been done in laboratory studies, e.g., Esteve-Gibert & Prieto, 2013; Pouw & Dixon, 2019b) to assess this relationship rather than explicitly measuring occurrence within a specific discrete boundary.

6.3.3. The effects of phrasal prosodic structure: The important role of phrasal position

As mentioned before, the two studies on gesture-speech temporal integration of the present thesis incorporated an assessment of the role of phrasal prosodic structure. The study in **Chapter 3** found that when a gesture occurs in an Accentual phrase (AP) where there are two potential anchoring sites, that is an initial accent (IA) and a final accent (FA), they tend to align with the IA more than the FA. In other words, in French, there seems to be a clear left-edge marking effect. Similarly, a "left-edge" effect was found in the results of the study in **Chapter 4,** showing that in English, gestures had a tendency to occur with prenuclear pitch accent, which tend to mark the left edge most frequently.

It is particularly interesting to have found a left-edge effect in both languages. In French, the left edge marking effect was quite strong, with IA being marked significantly more than FA when both prosodic landmarks were present. In English, the effect was slightly weaker as it did not reach statistical significance. However, the tendency was indeed present, and taking into consideration that the

most common pattern was for ips to be produced with only one prenuclear pitch accent, in terms of frequency, the left-edge pattern surfaced quite often. The difference between the two languages may largely be attributed to the fact that APs can only contain a maximum of two pitch accents, whereas English may have multiple prenuclear pitch accents. As a result, gestures have more potential anchoring points with which to associate. The reason why gesture location in the two languages displays a general "left-edge marking" tendency may lie in interactions with relative prominence, structural effects, and pragmatic meaning.

An important result from the present thesis is that prosodic prominence as a stand-alone factor does not seem to explain the tendency for gestures to associate with the phrase-initial edge. The study in **Chapter 4** showed that in English, this relationship was not driven by prominence. Precisely, gestures associated with prenuclear pitch accents that were assessed as being equally strong to the strongest prominence or even being relatively weaker than another pitch accent in the ip nearly 80% of the time. Though the current thesis did not include an analysis of relative prominence for French, one study by Hualde et al., (2016) showed that in general, IAs in French are generally not perceived as more prominent than FAs.

Discarding a direct prominence effect, a potential explanation for the left-edge marking tendency may be structural in nature. Indeed, French exhibits a principle of bipolarization (e.g., Di Cristo, 2000; see **subsection 1.3.1**), where pitch accents mark the right edge

obligatorily, and when another pitch accent is realized, it occurs on one of the first syllables of the AP. Similarly, English has been shown to follow the same principle to avoid stress clash (e.g., Shattuck-Hufnagel et al., 1994). In English, it has been specifically hypothesized that this effect may be "attention-getting" to raise listener awareness of the beginning of a new prosodic phrase (Bolinger, 1985). Thus, gestures may be functioning similarly, regularly marking the left edge to structurally mark the onset of a new prosodic phrase. A final explanation may be due to pragmatic factors. In French, IAs are optional and often convey an emphatic or pragmatic meaning. Therefore, speakers may choose to reinforce such meanings with gestures. However, this argument needs to be empirically tested in future studies taking both discourse genre as well as differences in IA realization into account. Regarding discourse genre, TED Talks represent didactic speech where speakers may be particularly expressive (Harrison, 2021) affecting prosodic strategies for communication. Moreover, the IA cannot be treated as a uniform phonological category – IA can be produced to convey pragmatic meaning but also for rhythmic purposes, and the two may be distinguished phonetically (see, e.g., Astésano, 2001; 2017). Future studies should distinguish rhythmic from pragmatic IA, and may also take relative prominence into account. Doing so may further elucidate (non-referential) gestures' role as a pragmatic vs. rhythmic marker in speech.

The aforementioned lack of effect of relative prosodic prominence in the attraction of gesture directly affects the assessment of the role of nuclear vs. prenuclear pitch accentuation in the gesture-speech

interface. Crucially, the results showing a left-edge effect in English represent an important contribution to the field by showing that neither the nuclear accent (e.g., Kendon, 1980) nor the "phonological peak" (as loosely defined by McNeill, 1992) systematically attract gesture, but prenuclear pitch accents (particularly those that are edge-initial) also act as strong gestural attractors, thus playing a key role in gesture-speech phonological synchrony.

In English, the study in **Chapter 5** offers some initial evidence regarding the pragmatic factors of gesture's attraction to prenuclear pitch accents. Namely, when referents are marked by prenuclear pitch accents, accessible referents are marked significantly more with an additional gestural cue than given or new referents. The role of prenuclear pitch accents in marking information status is unclear. While some studies have shown that the phonetic realization of prenuclear pitch accents is affected by IS (e.g., Kügler & Féry, 2008; Braun, 2006; see also Baumann, 2021), other studies claim they are produced merely to satisfy rhythmic constraints in speech (e.g., Calhoun, 2010b). It seems that if prenuclear pitch accents are produced to satisfy rhythmic constraints, gesture complements the pragmatic function of marking ISR by specifically associating with accessible information. From a first perspective, it seems that aspects of pragmatic meaning may guide which accents receive gesture and which ones do not. Additionally, these results show how gestures can continue contributing pragmatic meaning when the meaning conveyed by pitch accentuation is "less reliable." All in all, more work is needed to further disentangle the different

factors that lead to the “left-edge marking” tendency that this thesis has unveiled.

Summarizing the contributions of the current thesis to the phonological synchrony rule, our results highlight the key role of phrasal prosodic structure on gesture production and the necessity to take into account the specific language’s prosodic structure when assessing crosslinguistic alignment patterns. The results of the thesis show that not only prosodic prominence, but also prosodic phrasing (i.e., phrasal position) is key in the production patterns of co-speech gestures. Two main findings were reported in this regard. First, the studies in both French and English found a tendency for individual gesture strokes to mark the left edge of the prosodic phrase when two or more pitch accents are present (potentially answering why not all pitch accents receive a gesture). Second, the French study showed how even though RGGs do not seem to closely correspond to rhythmic prominence in speech (potentially answering why not all gestures associate with prominence in speech), it indeed remains influenced by the length of the basic AP prosodic structure domains in French.

6.3.4. The effects of phrasal prosodic structure: Phrasal duration predicts distance between RGG apexes

The study in **Chapter 3** investigated the temporal alignment patterns of RGGs. In this chapter we wanted to investigate whether the strokes of gestures belonging to RGGs are constrained temporally to associate with pitch accented syllables or rather their

temporal association patterns are independent of prosodic structure. Previous studies on this aspect have largely shown how beat-like groups of gesture (RGGs) are loosely associated with rhythmic prominence in speech. Specifically, they have reported that at least one gesture within the RGG associates with the nuclear pitch accent, and the others span out in both directions in a rhythmic fashion, where individual strokes within the RGG may or may not coincide with pitch accented syllables, suggesting that they follow their own tempo independent of pitch accentuation (McClave, 1994; see also Loehr, 2007). The results from the study in **Chapter 3** reinforced the findings in the previous literature, in that in French, RGGs did not correspond clearly with subsequent pitch accents. We similarly noted that sometimes subsequent gestures within a RGG would halve or double in the distance between each other, and importantly, this did not occur as a function of the number of pitch accents present within the AP. In other words, sometimes APs would contain two RGG apexes but only one pitch accent, and vice-versa. Thus, it seems that when groups of gestures are produced in a beat-like rhythmic fashion, pitch accentuation does not seem to be a guiding factor in their production. This finding along with those from the previous literature sheds some light as to why certain gestures may be produced without pitch accentuation, as posed in **subsection 6.3**. Importantly, the study in the current thesis is the first to assess whether phrasing may play a role in the production of RGGs. It found that indeed the duration of the AP was a significant predictor of the interval between RGG apexes. This finding is important as the previously mentioned studies suggest that the

rhythmic production of subsequent gestures is largely independent of rhythmic prominence in speech. This is the first study to find that while some gesture strokes (e.g., those within RGGs) are not dependent on pitch accentuation, they remain sensitive to prosodic structure. It further consolidates that close relationship exists between gestures and prosodic structure, not only in terms of pitch accentuation, but also in more complex ways through rhythm and phrasing.

These findings integrate quite nicely with more recent descriptions of rhythm as a multi-level phenomenon (e.g., Pouw et al., 2021). The Dynamic Systems Theory (McNeill, 1992, 2005; Iverson & Thelen, 1999; Rusiewicz et al., 2014; Pouw & Dixon, 2019a), assumes that rhythmic behavior is oscillatory in nature, and multiple oscillators may couple or entrain. It is proposed here that the "gesture oscillator" may decouple and subsequently entrain to multiple "speech oscillators" – that is, it may go from entraining to rhythms produced via pitch accents, to a syllabic rhythm and even to a phrasal rhythm. Future studies could use techniques to decompose the "speech oscillator" into its various components and assess entrainment at multiple levels to better understand the relationship between rhythm in speech and gesture, or even other non-linguistic rhythmic behaviors (e.g., breathing) to better understand the macrostructure of human rhythmic behavior. Importantly, there remain aspects of gesture-speech alignment that are independent of prosodic prominence (e.g., RGGs) which need further investigation.

6.4. Refining the pragmatic synchrony rule

The pragmatic synchrony rule claims that gesture and speech convey are coherent in conveying pragmatic meaning. While the same has been said about semantic meaning in the phonological synchrony rule (i.e., gesture and speech express the same semantic meaning), recent studies have shown that the semantic relationship is indeed much more complex. For example, referential gestures may indeed be completely redundant with speech, but they may also convey supplemental information that is not present in speech (see **subsection 1.1.** of the current thesis). The results from **Chapter 5** in the thesis have indeed shown that the pragmatic relationship between speech and gesture share similar complexities, at least in terms of the multimodal marking of ISR.

6.4.1. Multimodal cues to ISR

The results from the study in **Chapter 5** further advance our knowledge as to how gesture and pitch accentuation both function jointly to mark ISR. Specifically, our results showed for the first time that English speakers tend to mark new and accessible referents in academic discourses by both multimodal cues, while given referents are multimodally marked significantly less (i.e., they are more likely to be deaccented and be produced without a gesture). This is the first time that both the prosodic marking and the gesture marking of ISR constituents are jointly analyzed. In general, the findings are largely in line with previous studies in both the field of prosody (e.g., Im et al., 2018) where new and

accessible referents tend to receive pitch accentuation across languages, and only given referents tend to be deaccented, and the field of gesture (e.g., Debreslioska & Gullberg, 2019), where gestures tend to co-occur with new referents more often than given ones. Within the gesture field, the only study to our knowledge that accounts for the same levels of ISR and that obtains different results is that of Debreslioska & Gullberg (2020), where they found that gestures associate more with accessible than with new referents. The authors hypothesize that this may be due to the linguistic encoding of accessible referents with definite nominals (much like given referents), and thus gesture is being used as a strategy by the speaker to signal to the addressee that even though the form of the referent appears to be given, it is in fact not a given referent and should rather be treated as new. The difference in results may also be related to the different methods employed in the studies. In their study, they used a narrative retelling task carried out in a lab, whereas the current thesis used TED Talks, which may be a much more expressive genre of speech.

Interestingly, the results in **Chapter 5** showed that when referents were not marked by both multimodal cues, the next most common cue to mark ISR was pitch accentuation without gesture, while gesture-only marking of ISR appeared to be the least common. This suggests that gesture largely reinforces pitch accentuation to mark ISR. Furthermore, neither gesture type nor pitch accent type showed a probabilistic relationship with ISR. Though these results differ from previous studies on the prosodic marking of ISR (e.g., Im et al., 2018), a recent study by Baumann et al. (2021) was unable to

find any clear relationship between categorical pitch accent types and ISR. In their database on German read speech, referents in topic constituents were regularly produced with a rising pitch accent, regardless of their information status. In the current database, relative prominence showed to be a more robust marker than pitch accent type. The discrepancy between pitch accent type and relative prominence can be explained by methodological factors. Specifically, a paradigmatic approach was taken in the analysis of pitch accent type, looking at referents and their pitch accentuation across the entire database. It thus did not take into account the relationship between pitch accent types on a syntagmatic level. However, the assessment of relative prominence was indeed instantiated at the syntagmatic level, centering on relative prominence within the IP. Thus, perhaps by focusing on local relationships within prosodic phrases, clearer relationships between the encoding of pitch accent type and information status could come to light.

Finally, as previously mentioned, the results of **Chapter 5** showed that gesture production continues to be sensitive to information status when referents are marked by prenuclear pitch accents. Specifically, even though prenuclear pitch accentuation has been considered less stable for the marking of ISR (as per Calhoun, 2010a), gesture continues to mark accessible and new referents, and specifically marks accessible referents significantly more than new referents. This set of results is indeed in line with the study by Debreslioska & Gullberg (2020b) who found gesture to associate more with accessible referents over new referents. The authors

hypothesize that the morphosyntactic form of the accessible referent is ambiguous (as it is realized as a definite nominal expression much like a given referent). Thus, gesture is employed as a communicative strategy to signal to the addressee that the accessible referent should be treated as something relatively new. The results of the study in **Chapter 5** could be seen as parallel to the hypothesis by Debreslioka & Gullberg (2020a). namely that when a referent is in prenuclear position (a position often associated with given referents, and where pitch accentuation may be considered a more ambiguous marker of ISR), the gesture may be a disambiguating cue, associating with the accessible referent to indicate to speakers that it is indeed to be treated as a newer referent.

Summarizing the contributions of the current thesis to the pragmatic synchrony rule, the findings reported in **Chapter 5** have shown clear evidence that manual gestures are sensitive to the ISR distinction. Moreover, they are not directly parasitic on prosodic structure. That is, not only do speech and gesture convey the same pragmatic meaning, but that speech and gesture share a complex relationship in conveying pragmatic meaning, in that they may complement one another if one meaning-making strategy is less stable for other pragmatic (rhythmic) purposes. Importantly, the study focused on the information status of referents. Further work may wish to assess interactions at various levels of IS, or even assess other pragmatic aspects of speech (e.g., different speech acts). Moreover, it specifically investigated the production of gestures and pitch accentuation, and it would be important to

understand the effects of multimodal cues to IS for listeners. Future studies may want to investigate the perception and subsequent interpretation of these multimodal cues via online experimental methodologies such as Electroencephalography (e.g., Baumann & Schumacher, 2011; 2020) or eye-tracking (e.g. Braun & Biezma, 2019). Only then will we better understand the real value of multimodally marking information status for the listener.

6.5. Future work for M3D

M3D lays the groundwork for a full-fledged labeling system that accounts for a wide range of aspects of multimodal communication. However, the current thesis (and consequently, the development of M3D) has largely focused on the referential/non-referential distinction, its interaction with prosodic structure, and pragmatic function specifically in terms of ISR. More work is needed to develop other aspects of M3D which have remained largely theoretical as they have not been implemented in actual multimodal corpora. For example, while the M3D labeling guidelines offer a proposal for the annotation of manual gestures, head movements, and "other" articulators, the system has only been fully implemented for manual gestures. Though initial versions of M3D have indeed been applied to the Audiovisual corpus of Catalan children's narrative discourse development (Vilà-Giménez et al., 2022), which annotated head movements and other articulators, the guidelines could be further developed and standardized for such types of communicative movement.

Another aspect that future research should address is the annotation of the pragmatic functions of gesture. While M3D identifies several pragmatic domains and functions based on the literature, the current thesis has made headway specifically in the marking of ISR. M3D will be strengthened as more research assesses the different potential pragmatic domains that have been proposed thus far, potentially identifying additional functions not present, or using a data-driven method for identifying different pragmatic meanings that can be conveyed gesturally. Research in these areas may help develop more standard coding procedures, which would be a key next step for the field of gesture research. Finally, the M3D-TED corpora with which M3D was implemented represents but one specific discourse genre (academic-style speech). Though we have argued that TED Talks represent natural speech, it has indeed been rehearsed and is performed in a specific context (e.g., under time constraints). Therefore, M3D would be greatly strengthened by applying it to other genres of discourse (e.g., spontaneous dyadic conversation).

Given the abovementioned limitations, M3D represents an ongoing, long term project. It represents an important advance for the field of gesture research, namely by offering detailed guidelines for labelers to follow and will soon offer training material so that labelers can not only read about, but practice and truly learn how to annotate multimodal corpora in a standard way. This will allow for future studies that employ M3D to be directly comparable and will foster more interdisciplinary research across the field of gesture, prosody, semantics, and pragmatics.

6.6. Final conclusions

All in all, the four empirical studies in this thesis contribute to our understanding of the prosodic and pragmatic properties of gestures across languages. Specifically, they advance our knowledge about the phonological and pragmatic synchrony rules. Regarding the properties of gesture-speech alignment, the results demonstrate that gesture temporal association properties are not based on prosodic prominence alone, but also on positional properties within phrasal prosodic structure. Importantly, this positional effect was found in two typologically different languages in terms of speech prosody, French and English. Regarding the pragmatic properties of gesture as ISR markers, the results have shown that even though prosodic and gesture prominence work in an integrated fashion, gesture is not directly parasitic on prosodic structure and can display ISR marking functions. Importantly, these empirical findings lend support to M3D as a valid approach to the study of multimodal language, focusing on interaction between modes of communication and where gestures are assessed across three largely independent, non-mutually exclusive dimensions, namely the form dimension, the prosodic dimension, and the meaning dimension. Ultimately, M3D offers an opportunity to advance the field towards adopting a standardized, multidisciplinary approach to the study of gestures.

References

Abner, N., Cooperrider, K. & Goldin-Meadow, S. (2015). Gesture for Linguists: A Handy Primer. *Language and Linguistics Compass, 9*(11), 437–451. https://doi.org/10.1111/lnc3.12168

Aguilar, L., De-la-Mota, C., & Prieto, P. (2011). Cat_ToBI Training Materials. http://prosodia.upf.edu/cat_tobi/

Alexanderson, S., House, D. & Beskow, J. (2013). Aspects of co-occurring syllables and head nods in spontaneous dialogue. *Proceedings of the 12th international conference on Auditory-Visual Speech Processing (AVSP2013)*, 169-172, https://www.isca-speech.org/archive/avsp_2013/alexanderson13_avsp.html

Alibali, M. W., & Goldin-Meadow, S. (1993). Transitions in learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468–523. https://doi.org/10.1006/cogp.1993.1012

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10*(1), 3–28. https://doi.org/10.1075/gest.10.1.02ali

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science, 22*(9), 1138–1144. https://doi.org/10.1177/0956797611417722

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation, 41*, 273–287.

Ambrazaitis, G. (2009). Nuclear intonation in Swedish: Evidence from Experimental-Phonetic Studies and a Comparison with German. *Travaux de L'institut de Linguistique de Lund, 49*. Lund University Press.

Ambrazaitis, G. & House, D. (2016). Multimodal levels of prominence: the use of eyebrows and head beats to convey information structure in Swedish news reading. *7th Conference of the International Society for Gesture Studies*. 319–319.

Ambrazaitis, G. & House, D. (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication, 95*, 100–113. https://doi.org/10.1016/j.specom.2017.08.008

Ambrazaitis, G. & House, D. (2022). Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters. *Laboratory Phonology, 24*(1). https://doi.org/10.16995/labphon.6430

Ambrazaitis, G., Zellers, M. & House, D. (2020). Compounds in interaction: patterns of synchronization between manual

gestures and lexically stressed syllables in spontaneous Swedish. *The 7th Gesture and Speech in Interaction (GESPIN2020)*. Stockholm, Sweden. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1539106

Anderson, C. (2016). *TED Talks: The official TED guide to public speaking: Tips and tricks for giving unforgettable speeches and presentations*. Hachette UK.

Andric, M., Solodkin, A., Buccino, G., Goldin-Meadow, S., Rizzolatti, G., & Small, S. (2013). Brain function overlaps when people observe emblems, speech, and grasping. *Neuropsychologia*, *51*, 1619–1629. http://doi.org/10.1016/j.neuropsychologia.2013.03.022

Arnhold, A., Chen, A. & Järvikivi, J. (2016). Acquiring complex focus-marking: Finnish 4-to 5-year-olds use prosody and word order in interaction. *Frontiers in Psychology*, *7*, 1886. https://doi.org/10.3389/fpsyg.2016.01886

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*, 555–596. http://doi.org/10.1162/coli.07-034-R2

Arvaniti, A. (2022). The Autosegmental-Metrical model of intonational phonology. In J. Barnes & S. Shattuck-Hufnagel (Eds), *Prosodic Theory and Practice* (pp. 26-63). The MIT Press.

Astésano, C. (2001). *Rythme et Accentuation en Français: Invariance et Variabilité Stylistique*. L'Harmattan.

Astésano, C. (2017). *Le statut de l'Accent Initial dans la phonologie prosodique du français: enjeux descriptifs et psycholinguistiques* [Habilitat]. Université de Toulouse - Jean Jaurès.

Astésano, C. (2022). *De la supramodalité du rythme — Implications pour la description prosodique, la remédiation linguistique et l'apprentissage des langues* [Oral presentation]. Journées d'Etudes sur la Parole, 2022. Noirmoutier, France.

Astésano, C., Bard, E. G. & Turk, A. (2007). Structural influences on initial accent placement in French. *Language and Speech*, 50(3), 423–446. https://doi.org/10.1177/00238309070500030501

Austin, E. E. & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of Experimental Child Psychology, 122*, 92–103. https://doi.org/10.1016/j.jecp.2013.12.008

Ayers, G. (1996). *Nuclear accent types and prominence: some psycholinguistic experiments* [Unpublished Doctoral Dissertation]. Ohio State University.

Baills, F., Rohrer, P. L., & Prieto, P. (2022). Le geste et la voix pour enseigner la prononciation en langue étrangère. *Mélanges CRAPEL*, *43*(1), 157–184.

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, *41*(1), 33–58. https://doi.org/10.1017/S0272263118000074

Barros, C. A. (2021). *A relação entre unidades gestuais e quebras prosódicas: o caso da unidade informacional Parentético* [Unpublished Master's Dissertation]. Universidade Federal de Minas Gerais.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Baumann, S. (2006). Information structure and prosody: Linguistic categories for spoken language annotation. *Methods in Empirical Prosody Research*, *3*, 153–180. https://doi.org/10.1515/9783110914641.153

Baumann, S. (2016). Second Occurrence Focus. In C. Féry & S. Ishihara (Eds.), *Oxford Handbook of Information Structure* (pp. 483–502). Oxford University Press.

Baumann, S., Becker, J., Grice, M. & Mücke, D. (2007). Tonal and articulatory marking of focus in German. *Proceedings of the 16th International Congress of Phonetic Sciences,* 1029–1032.

Baumann, S. & Grice, M. (2006). The intonation of accessibility. *Journal of Pragmatics, 38*(10), 1636–1657. https://doi.org/10.1016/j.pragma.2005.03.017

Baumann, S., Grice, M. & Steindamm, S. (2006). Prosodic marking of focus domains-categorical or gradient. *Proceedings of Speech Prosody 2006*, 301–304.

Baumann, S. & Kügler, F. (2015). Prosody and information status in typological perspective-Introduction to the Special Issue. *Lingua*, *165*, 179–182. https://doi.org/10.1016/j.lingua.2015.08.001

Baumann, S., Mertens, J., & Kalbertodt, J. (2019). Informativeness and speaking style affect the realization of nuclear and prenuclear accents in German. In S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1580-1584). Australasian Speech Science and Technology Association Inc.

Baumann, S., Mertens, J. & Kalbertodt, J. (2021). The influence of informativeness on the prosody of sentence topics. Glossa: A *Journal of General Linguistics, 6*(1), 1-28. https://doi.org/10.16995/glossa.5871

Baumann, S. & Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and Meaning*, *25*, 119–162. https://doi.org/10.1515/9783110261790.119

Baumann, S. & Riester, A. (2013). Coreference, lexical givenness and prosody in German. *Lingua*, *136*, 16–37. https://doi.org/10.1016/j.lingua.2013.07.012

Baumann, S. & Röhr, C. T. (2015). The perceptual prominence of pitch accent types in German. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the International Congress of Phonetic Sciences*. IPA Public Archive.

Baumann, S. & Schumacher, P. B. (2012). (De-)accentuation and the process of information status: evidence from event-related brain potentials. *Language and Speech*, *55*(3), 361–381. https://doi.org/10.1177/0023830911422184

Baumann, S. & Schumacher, P. (2020). The incremental processing of focus, givenness and prosodic prominence. *Glossa: a journal of general linguistics*, *5*(1), 6. 1-30. https://doi.org/10.5334/gjgl.914

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse processes*, *15*, 469–489. http://doi.org/10.1080/01638539209544823

Beaver, D. I., Clark, B., Flemming, E. S., Jaeger, T. F. & Wolters, M. (2007). When semantics meets phonetics: Acoustical studies of second-occurrence focus. *Language, 83*(2), 245–276. http://doi.org/10.1353/lan.2007.0053

Berger, S. & Zellers, M. (2022). Multimodal prominence marking in semi-spontaneous YouTube monologues: the interaction of intonation and eyebrow movements. *Frontiers in Communication,* 132. http://doi.org/10.3389/fcomm.2022.903015

Bergmann, K., Aksu, V. & Kopp, S. (2011). The relation of speech and gestures: temporal synchrony follows semantic synchrony. In S. Kopp, K. Rohlfing, J.de Ruiter, P. Wagner, M. Karpinski (Eds.), *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction*.

Biau, E., Fernández, L. M., Holle, H., Avila, C. & Soto-Faraco, S. (2016). Hand gestures as visual prosody: BOLD responses to audio-visual alignment are modulated by the communicative nature of the stimuli. *NeuroImage, 132*, 129–137. https://doi.org/10.1016/j.neuroimage.2016.02.018

Blache, P., Bertrand, R., Ferré, G., Pallaud, B., Prévot, L., & Rauzy, S. (2017). The corpus of interactional data: A large multimodal annotated resource. In N. Ide and J. Pustejovsky (Eds.), *Handbook of Linguistic Annotation* (pp. 1323-1356). Springer.

Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.14.

Bolinger, D. (1985). Two views of accent. *Journal of Linguistics, 21*(1), 79–123. https://www.jstor.org/stable/4175764

Bolly, C. T. (2016*). CorpAGEst Annotation Manual (II. Speech Annotation Guidelines)*. Available from https://corpagest.wordpress.com/working-papers/

Bolly, C. T., & Boutet, D. (2018). The multimodal CorpAGEst corpus: Keeping an eye on pragmatic competence in later life. *Corpora, 13*, 279–317. https://doi.org/10.3366/cor.2018.0151

Bressem, J. (2013). A linguistic perspective on the notation of form features in gestures. In C. Müller, E. Fricke, A. Cienki, D. McNeill, S. Ladewig, & S. Tessendorf (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)* 38/1 (pp. 1079–1098). De Gruyter. http://doi.org/10.1515/9783110261318.1079

Bressem, J., Ladewig, S. H., & Müller, C. (2013). Linguistic Annotation System for Gestures. In C. Müller, E. Fricke, A. Cienki, D. McNeill, S. Ladewig, & S. Tessendorf (Eds.), *Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK), 38/1*, (pp. 1098–1124). De Gruyter.

Braun, B. (2006). Phonetics and phonology of thematic contrast in German. *Language and Speech*, *49*(4), 451–493. https://doi.org/10.1177/00238309060490040201

Braun, B. & Biezma, M. (2019). Prenuclear L∗+ H activates alternatives for the accented word. *Frontiers in Psychology, 10*, 1993. http://doi.org/10.3389/fpsyg.2019.01993

Breen, M., Fedorenko, E., Wagner, M. & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes, 25*(7-9), 1044–1098. http://dx.doi.org/10.1080/01690965.2010.504378

Broaders, S., Cook, S.W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture reveals implicit knowledge and leads to learning. *Journal of Experimental Psychology, 136*(4), 539–550. https://doi.org/10.1037/0096-3445.136.4.539

Brown, G. (1983). Prosodic structure and the given/new distinction. In A. Cutler & D. A. Ladd (Eds.), *Prosody: Models and measurements* (pp. 67–77). Springer.

Burchardt, L. S. & Knörnschild, M. (2020). Comparison of methods for rhythm analysis of complex animals' acoustic signals. PLoS *Computational Biology, 16*(4), e1007755. https://doi.org/10.1371/journal.pcbi.1007755

Büring, D. (1997). *The meaning of topic and focus*. Routledge.

Büring, D. (2003). On D-trees, beans, and B-accents. *Linguistics and Philosophy*, *26*(5), 511–545. https://doi.org/10.1023/A:1025887707652

Büring D. (2007). Intonation, Semantics and Information Structure," In G. Ramchand & C. Reiss (Eds.), The Oxford Handbook of Linguistic Interfaces (pp. 445-474). Oxford University Press.

Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one- to two-word speech: When hand and mouth come together. In D. McNeill (Ed.), *Language and gesture* (pp. 235–258). Cambridge University Press.

Butterworth, B., Swallow, J., & Grimston, M. (1981). Gestures and Lexical Processes in Jargonaphasia. In J.W. Brown (Ed.), *Jargonaphasia* (pp. 113–124). Academic Press. http://doi.org/10.1016/C2013-0-07219-2

Calhoun, S. (2009). What makes a word contrastive? Prosodic, semantic and pragmatic perspectives. In D. Barth-Weingarten, N. Dehé, and A. Wichmann (Eds.), *Where prosody meets pragmatics* (pp. 53–78). Emerald Group Publishing.

Calhoun, S. (2010a). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, *25*(7-9), 1099–1140. https://doi.org/10.1080/01690965.2010.491682

Calhoun, S. (2010b). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language, 86*(1), 1–42. https://www.jstor.org/stable/40666298

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics, 40*(2), 329–349. https://doi.org/10.1016/j.wocn.2011.12.001

Cangemi, F. & Grice, M. (2016). The Importance of a Distributional Approach to Categoriality in Autosegmental-Metrical Accounts of Intonation. *Laboratory Phonology, 7*(1), 9. http://dx.doi.org/10.5334/labphon.28

Cantalini, G. & Moneglia, M. (2020). The annotation of gesture and gesture/prosody synchronization in multimodal speech corpora. *Journal of Speech Sciences, 9*, 07–30. https://doi.org/10.20396/joss.v9i00.14956

Capirci, O., Iverson, J.M., Pizzuto, E., & Volterra, V. (1996). Communicative gestures during the transition to two-word speech. *Journal of Child Language, 23*, 645–673. https://doi.org/10.1017/S0305000900008989

Cavé, C., Guaïtella, I., Bertrand, R., Santi, S., Harlay, F. & Espesser, R. (1996). About the relationship between eyebrow movements and Fo variations. *Proceedings of the Fourth International Conference on Spoken Language Processing*, *4*, 2175–2178. http://doi.org/10.1109/ICSLP.1996.607235

Chafe, W. L. (1974). Language and consciousness. *Language*, *50*(1), 111–133. https://doi.org/10.2307/412014

Cho, T. (2016). Prosodic boundary strengthening in the phonetics-prosody interface. *Language and Linguistics Compass, 10*(3), 120–141. https://doi.org/10.1111/lnc3.12178

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*, 43–71. http://doi.org/10.1016/0010-0277(86)90053-3

Cienki, A., & Müller, C. (2008). Metaphor, gesture, and thought. In R. W. Gibbs, Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 483–501). Cambridge University Press. https://doi.org/10.1017/CBO9780511816802.029

Clark, H. H. (1975). Bridging. Theoretical Issues in Natural Language Processing. In B. L. Nash-Webber & R. Schank (Eds.), *Proceedings of the 1975 workshop on theoretical issues in natural language processing* (pp. 169–174). Association for Computational Linguistics.

Clark, H. H. (1977). Inferences in comprehension. In D. LaBerge and S. J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 243–263). Lawrence Erlbaum Associates.

Cohen, R. L. & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology, 4*(2), 113–139. https://doi.org/10.1080/09541449208406246

Cole, J. & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology, 7*(1). http://dx.doi.org/10.5334/labphon.29

Cook, S. W., Mitchell, Z. & Goldin-Meadow, S. (2008). Gesturing makes learning last. Cognition, 106(2), 1047–1058. https://doi.org/10.1016%2Fj.cognition.2007.04.010

Cook, S. W., Yip, T. K. & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes, 27*(4), 594–610. https://doi.org/10.1080/01690965.2011.567074

Cooperrider, K., Abner, N. & Goldin-Meadow, S. (2018). The Palm-Up Puzzle: Meanings and Origins of a Widespread Form in Gesture and Sign. *Frontiers in Communication, 3*. https://doi.org/10.3389/fcomm.2018.00023

Couper-Kuhlen, E. & Barth-Weingarten, D. (2011). *A system for transcribing talk-in-interaction: GAT 2: English translation and adaptation of Selting, Margret et al.,*

Gesprächsanalytisches Transkriptionssystem 2. Gesprächsforschung 12, 1–51.

Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

Crystal, D. & Davy, D. (1969). *Investigating English style*. Indiana University Press.

Dargue, N. & Sweller, N. (2020). Learning stories through gesture: Gesture's effects on child and adult narrative comprehension. *Educational Psychology Review, 32*(1), 249–276. https://doi.org/10.1007/s10648-019-09505-0

Debreslioska, S., & Gullberg, M. (2019). Discourse reference is bimodal: how information status in speech interacts with presence and viewpoint of gestures. *Discourse Processes, 56*, 41–60. http://doi.org/10.1080/0163853X.2017.1351909

Debreslioska, S., & Gullberg, M. (2020a). The semantic content of gestures varies with definiteness, information status and clause structure. *Journal of Pragmatics, 168*, 36–52. http://doi.org/10.1016/j.pragma.2020.06.005

Debreslioska, S., & Gullberg, M. (2020b). What's New? Gestures accompany inferable rather than brand-new referents in discourse. *Frontiers in Psychology, 11*, 1935. http://doi.org/10.3389/fpsyg.2020.01935

Debreslioska, S. & Gullberg, M. (in press). Information status predicts the incidence of gesture in discourse-an experimental study. *Discourse Processes*. https://doi.org/10.1080/0163853X.2022.2085476

Debreslioska, S., Özyürek, A., Gullberg, M., & Perniss, P. (2013). Gestural viewpoint signals referent accessibility. *Discourse Processes*, *50*, 431–456. http://doi.org/10.1080/0163853X.2013.824286

Delais-Roussarie, E. (1995). *Pour une approche parallèle de la structure prosodique: étude de l'organisation prosodique et rythmique de la phrase française* [Unpublished doctoral Dissertation]. Université de Toulouse - Le Mirail.

Delais-Roussarie, E. (1996). Phonological Phrasing and Accentuation in French. In M. Nespor & N. Smith (Eds.), *Dam Phonology : HIL Phonology Paper II* (pp. 1-38). Holland Academic Graphics.

Delais-Roussarie, E. & Di Cristo, A. (2021). XIX-4 L'accentuation. In Abeillé A. & Godard D. (Eds.), *La grande Grammaire du français - GGF: Volume 2* (pp. 2126–2139). Actes Sud. https://halshs.archives-ouvertes.fr/halshs-00748395

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R. & Yoo, H.-Y. (2015). Intonational phonology of French: Developing a ToBI system

for French. In S. Frota & P. Prieto (Eds.), *Intonation in romance* (pp. 63-100). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685332.003.0003

Dell, F. (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst, and J.-R. Vergnaud (Eds.), *Forme sonore du langage* (pp. 65–122). Hermann.

Demir, Ö.E., Levine, S.C., Goldin-Meadow, S. (2015). A tale of two hands: Children's early gesture use in narrative production predicts later narrative structure in speech. *Journal of Child Language, 42*, 662–681. https://doi.org/10.1017/s0305000914000415

Dettori, J. R., & Norvell, D. C. (2020). Kappa and beyond: Is there agreement? *Global Spine Journal, 10*, 499–501. http://doi.org/10.1177/2192568220911648

Di Cristo, A. (1998). Intonation in French. In D. J. Hirst & A. Di Cristo (Eds.), *Intonation systems: A survey of twenty languages* (pp. 195-218). Cambridge University Press.

Di Cristo, A. (1999). Vers une modélisation de l'accentuation du français: première partie. *Journal of French Language Studies, 9*(2), 143–179. https://doi.org/10.1017/S0959269500004671

Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français (seconde partie). *Journal of French Language*

Studies, 10(1), 27–44. https://doi.org/10.1017/S0959269500000120

Di Cristo, A. (2016). *Les musiques du français parlé. Essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain. Études de linguistique française (vol. 1)*. De Gruyter Mouton.

Dilley, L. & Brown, M. (2005). *The RaP (Rhythm and Pitch) Labeling System, v. 1.0* [Unpublished Manuscript]. https://tedlab.mit.edu/tedlab_website/RaP%20System/RaP_Labeling_Guide_v1.0.pdf

Dimitrova, D., Chu, M., Wang, L., Özyürek, A. & Hagoort, P. (2016). Beat that word: How listeners integrate beat gesture and focus in multimodal speech discourse. *Journal of Cognitive Neuroscience, 28*(9), 1255–1269. https://doi.org/10.1162/jocn_a_00963

Dohen, M., Løevenbruck, H. & Hill, H. C. (2006). Visual correlates of prosodic contrastive focus in French: description and inter-speaker variability. In R. Hoffmann & H. Mixdorff (Eds.), *Speech Prosody 2006* (pp. 221-224). ISCA Archive.

Driskell, J. E. & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. *Human Factors, 45*(3), 445–454. https://doi.org/10.1518/hfes.45.3.445.27258

Duboisdindien, G. (2019). *Analyse multimodale des marqueurs pragmatiques au sein du vieillissement langagier en situation de Trouble Cognitif Léger* [Unpublished Doctoral Dissertation]. Université Paris - Nanterre.

Ebert, C., Evert, S. & Wilmes, K. (2011). Focus marking via gestures. In I. Reich, E. Horch, & D. Pauly (Eds.), *Proceedings of Sinn Und Bedeutung 15* (pp. 193–208).

Efron, D. (1941). *Gesture and environment*. King's Crown Press.

Ekman, P. & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*(1), 49–98. https://doi.org/10.1515/9783110880021.57

Elvira García, W. (2017). Create pictures with tiers v.4.4. [Praat script]. Retrieved from http://stel.ub.edu/labfon/en/praat-scripts

Esposito, A., Esposito, D., Refice, M., Savino, M. & Shattuck-Hufnagel, S. (2007). A preliminary investigation of the relationship between gestures and prosody in Italian. In A. Esposito, M. Bratanić, E. Keller, and M. Marinaro (Eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue* (pp. 65 – 74). IOS Press.

Esteve-Gibert, N., Borràs-Comes, J., Asor, E., Swerts, M. & Prieto, P. (2017). The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical*

Society of America, 141(6), 4727–4739. https://doi.org/10.1121/1.4986649

Esteve-Gibert, N. & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research , 56*(3), 850–864. https://doi.org/10.1044/1092-4388(2012/12-0049)

Ferré, G. (2010). Timing relationships between speech and co-verbal gestures in spontaneous French. *Language Resources and Evaluation, Workshop on Multimodal Corpora, 6*, 86–91.

Ferré, G. (2011). Functions of three open-palm hand gestures. *Journal Multimodal Communication, 1*(1), 5–20. https://doi.org/10.1515/mc-2012-0002

Ferré, G. (2014). A multimodal approach to markedness in spoken French. *Speech Communication, 57*, 268–282. https://doi.org/10.1016/j.specom.2013.06.002

Féry, C. (2017). *Intonation and prosodic structure*. Cambridge University Press.

Féry, C. & Samek-Lodovici, V. (2006). Focus projection and prosodic prominence in nested foci. *Language, 82*(1), 131–150. https://www.jstor.org/stable/4490087

Féry, C. & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics, 36*(4), 680–703. https://doi.org/10.1016/j.wocn.2008.05.001

Feyereisen, P. (1998). Le rôle des gestes dans la mémorisation d'énoncés oraux. In S. Santi, I. Guaïtella, C. Cave et G. Konopczynski (Eds.), *Oralité et gestualité. Communication multimodale, interaction. Actes du colloque Orage 98* (pp. 355-360). L'Harmattan.

Firbas, J. (1971). On the concept of communicative dynamism in the theory of functional sentence perspective. *Philologica Pragensia, 8*, 135–144.

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication, 52*(6), 542–554. http://dx.doi.org/10.1016/j.specom.2009.12.003

Foraker, S. (2011). Gesture and discourse: how we use our hands to introduce and refer back. In G. Stam, M. Ishino, & R. Ashley (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 279-292). Benjamins. http://doi.org/10.1075/gs.4.26for

Fougeron, C. & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America, 101*(6), 3728–3740. https://doi.org/10.1121/1.418332

Freedman, N. & Hoffman, S. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills, 24*, 527-539. https://doi.org/10.2466/pms.1967.24.2.527

Frota, S., Oliveira, P. ,Cruz, M. & Vigário, M. (2015). P-ToBI: tools for the transcription of Portuguese prosody. http://labfon.letras.ulisboa.pt/InAPoP/P-ToBI/

Fung, H. S. H. & Mok, P. P. K. (2018). Temporal coordination between focus prosody and pointing gestures in Cantonese. *Journal of Phonetics, 71*, 113–125. https://doi.org/https://doi.org/10.1016/j.wocn.2018.07.006

Gerwing, J. & Bavelas, J. (2004). Linguistic influences on gesture's form. *Gesture, 4*(2), 157–195. https://doi.org/10.1075/gest.4.2.04ger

Givón, T. (1983). Topic continuity in discourse: an introduction. In Givon (Ed.), *Topic continuity in discourse: A quantitative cross-language study* (pp. 1-42). John Benjamins. http://doi.org/ 10.1075/tsl.3.01giv

Gilbert, E. (2009, February). *Your elusive creative genius*. [video]. TED Conferences. https://www.ted.com/talks/elizabeth_gilbert_your_elusive_creative_genius

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 85–107). Erlbaum.

Goldin-Meadow, S., Cook, S.W., & Mitchell, Z.A. (2009). Gesturing gives children new ideas about math. *Psychological Science, 20*(3), 267–272. https://doi.org/10.1111%2Fj.1467-9280.2009.02297.x

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics, 32*(10), 1489–1522. https://doi.org/10.1016/S0378-2166(99)00096-X

Götze, M., Weskott, T., Endriss, C., Fiedler, I., Hinterwimmer, S., Petrova, S., Schwarz, A., Skopeteas, S. & Stoel, R. (2007). *Information structure. Interdisciplinary Studies on Information Structure, 7*, 147–187.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology, 7*, 515-546. https://doi.org/10.1515/9783110197105.2.515

Graziano, M. & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology, 9*, 879. https://doi.org/10.3389/fpsyg.2018.00879

Graziano, M., Nicoladis, E. & Marentette, P. (2020). How Referential Gestures Align With Speech: Evidence From Monolingual and Bilingual Speakers. *Language Learning, 70*(1), 266–304. https://doi.org/10.1111/lang.12376

Guaïtella, I., Santi, S., Lagrue, B. & Cavé, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. *Language and Speech, 52*(2-3), 207–222. https://doi.org/10.1177%2F0023830909103167

Guellaï, B., Langus, A. & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology, 5*, 700. https://doi.org/10.3389/fpsyg.2014.00700

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish.* Lund University Press.

Gullberg, M. (2003). Gestures, referents, and anaphoric linkage in learner varieties. In C. Dimroth and M. Starren (Eds.), *Information Structure, linguistic Structure and the dynamics of language acquisition* (pp. 311-328). John Benjamins. http://doi.org/10.1075/sibil.26.15gul

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language Learning, 56*(1), 155–196. https://doi.org/10.1111/j.0023-8333.2006.00344.x

Gundel, J. K. (1996). Relevance theory meets the givenness hierarchy: an account of inferrables. In T. Fretheim & J. Gundel (Eds.), *Reference and referent accessibility* (pp. 141-153). John Benjamins.

Gunlogson, C. (2004). *True to form: Rising and falling declaratives as questions in English.* Routledge.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents.* Foris Publications.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In B. Bel and I. Marlien (Eds.), *Proceedings of the International Conference on Speech Prosody 2002* (pp. 47-57). Vientiane: International Speech Communication Association.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *The British Journal of Mathematical and Statistical Psychology, 61*, 29–48. http://doi.org/10.1348/ 000711006X126600

Habets, B., Kita, S., Shao, Z., Özyürek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech–gesture integration during comprehension. *Journal of Cognitive Neuroscience, 23*, 1845–1854. http://doi.org/10.1162/jocn.2010.21462

Halliday, M. A. (1967). Notes on transitivity and theme in English Part I. *Journal of Linguistics, 3*(1), 37–81. https://www.jstor.org/stable/4174950

Hardison, D. M. (2018). Visualizing the acoustic and gestural beats of emphasis in multimodal discourse: Theoretical and pedagogical implications. *Journal of Second Language Pronunciation, 4*(2), 232–259. https://doi.org/10.1075/jslp.17006.har

Harrison, S. (2021). Showing as sense-making in oral presentations: The speech-gesture-slide interplay in TED Talks by Professor Brian Cox. *Journal of English for Academic Purposes, 53*, 101002. https://doi.org/10.1016/j.jeap.2021.101002

Hirschberg, J. (1993). Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence, 63*(1-2), 305–340. https://doi.org/10.1016/0004-3702(93)90020-C

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication, 36*(1-2), 31-43. https://doi.org/10.1016/S0167-6393(01)00024-3

Hirst, D. (2007). A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation. In J. Trouvain and W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences XVI* (pp. 1233-1236). Pirrot GmbH.

Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for intonation. In M. Horne (Ed.), *Prosody: Theory and experiment* (pp. 51–87). Kluwer Academic Publishers.

Holle, H. & Rein, R. (2015). EasyDIAg: A tool for easy determination of interrater agreement. *Behavior research methods, 47*, 837–847. https://doi.org/10.3758/s13428-014-0506-7

Holler, J., Bavelas, J., Woods, J., Geiger, M. & Simons, L. (in press). Given-New Effects on the Duration of Gestures and of Words in Face-to-Face Dialogue. *Discourse Processes*, 1–27. https://doi.org/10.1080/0163853X.2022.2107859

Hualde, J. I., Cole, J., Smith, C., Eager, C., Mahrt, T. & Napoleão de Souza, R. (2016). The perception of phrasal prominence in English, Spanish and French conversational speech. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (Eds.), *Speech Prosody 2016* (pp. 459–463). ISCA Archive. http://doi.org/10.21437/SpeechProsody.2016-94

Hubbard, A. L., Wilson, S. M., Callan, D. E. & Dapretto, M. (2009). Giving speech a hand: gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, *30*(3), 1028–1037. https://doi.org/10.1002/hbm.20565

Hübscher, I., & Prieto, P. (2019). Gestural and prosodic development act as sister systems and jointly pave the way for children's sociopragmatic development. *Frontiers in Psychology, 10*, 1259. https://doi.org/10.3389%2Ffpsyg.2019.01259

Igualada A., Bosch, L., & Prieto, P. (2015). Language development at 18 months is related to communicative strategies at 12 months. *Infant Behavior and Development, 39*, 42-52. https://doi.org/10.1016/j.infbeh.2015.02.004

Im, S. & Baumann, S. (2020). Probabilistic relation between co-speech gestures, pitch accents and information status. *Proceedings of the Linguistic Society of America, 5*(1), 685. https://doi.org/10.3765/plsa.v5i1.4755

Im, S., Cole, J. & Baumann, S. (2018). The probabilistic relationship between pitch accents and information status in public speech. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński & D. Śledziński (Eds.), *Speech Prosody 2018* (pp. 508–511). ISCA Archive. http://doi.org/10.21437/SpeechProsody.2018-103

Iverson, J. M. & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6*, 11(12), 19–40.

Jankowski, L., Astésano, C. & Di Cristo, A. (1999). The initial rhythmic accent in French: Acoustic data and perceptual investigation. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *14th International Congress of Phonetic Sciences* (pp. 257–260). ICPhS Archive.

Jannedy, S. & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. *Interdisciplinary Studies on Information Structure*, 3, 199–244.

Jarmolowicz, E., Karpinski, M., Malisz, Z., & Szczyszek, M. (2007). Gesture, prosody and lexicon in task-oriented dialogues: multimedia corpus recording and labelling. In A. Esposito, M. Faundez-Zanuy, E. Keller, and M. Marinaro (Eds.), *Verbal and nonverbal communication behaviours* (pp. 99–110). Springer.

Johnson, S. (2010, July). *Where good ideas come from*. [video]. TED Conferences. https://www.ted.com/talks/steven_johnson_where_good_ideas_come_from/

Jun, S.-A. (2005). *Prosodic Typology: The phonology of intonation and phrasing*. Oxford University Press.

Jun, S.-A. (2014). *Prosodic Typology II: The phonology of intonation and phrasing*. Oxford University Press.

Jun, S.-A. & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation: Analysis, Modeling and Technology* (pp. 209–242). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-4317-2_10

Jun, S.-A. & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus, 14*(1), 147–172. https://doi.org/10.1515/prbs.2002.002

Karpiński, M., Jarmołowicz-Nowikow, E. & Malisz, Z. (2009). Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogues. *Speech and Language Technology, 11,* 113–122.

Kaufman, D. & Farinella, A. (2022). Gesture alignment in a "stressless" language. In T. Clark, J. Dussere & C. Ting (Eds.), Proceedings of the 28th Meeting of the Austronesian Formal Linguistics Association (pp. 29-46).

Keith, D. (2007). *A critical look at geoengineering against climate change.* [video]. TED Conferences. https://www.ted.com/talks/david_keith_a_critical_look_at_geoengineering_against_climate_change

Kelly, S. D., Kravitz, C. & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253–260. https://doi.org/10.1016/S0093-934X(03)00335-3

Kendon, A. (1972). Some relationships between body motion and speech: An analysis of an example. In A. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177-210). Pergamon Press. https://doi.org/10.1016/B978-0-08-015867-9.50013-7

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key (Ed.), *The relationship of verbal and nonverbal communication* (pp. 207–227). Mouton.

Kendon, A. (1982). The study of gesture: some observations on its history. *Recherches Sémiotiques/Semiotic Inquiry, 2*(1), 45–62. http://doi.org/10.1075/gest.16.2.01ken

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics, 23*, 247–279. http://doi.org/10.1016/0378-2166(94)00037-F

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kendon, A. (2017). Pragmatic functions of gestures. *Gesture, 16*, 157–175. http://doi.org/10.1075/gest.16.2.01ken.

Kita, S., Van Gijn, I. & Van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth & M. Fröhlich (Eds.),

Gesture and Sign Language in Human- Computer Interaction : International Gesture Workshop (pp. 23-35). Springer.

Krahmer, E. & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Krifka, M. (1998). Scope inversion under the rise–fall contour in German. *Linguistic Inquiry 29*, 75–112. https://www.jstor.org/stable/4179008

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica, 55*(3-4), 243–276. https://doi.org/10.1556/ALing.55.2008.3-4.2

Krivokapić, J., Tiede, M. K. & Tyrone, M. E. (2017). A Kinematic Study of Prosodic Structure in Articulatory and Manual Gestures: Results from a Novel Method of Data Collection. *Laboratory Phonology, 8*(1). https://doi.org/10.5334/labphon.75

Krivokapić, J., Tiede, M., Tyrone, M. E. & Goldenberg, D. (2016). Speech and manual gesture coordination in a pointing task. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (Eds.), *Speech Prosody 2016* pp.(1240-1244). ISCA Archive. http://doi.org/10.21437/SpeechProsody.2016-255

Kügler, F. & Calhoun, S. (2020). Prosodic encoding of information structure: A typological perspective. In C. Gussenhoven & A. Chen (Eds.), *Oxford handbook of language prosody* (pp. 454–467). Oxford University Press. https://doi.org/https://doi.org/10.1093/oxfordhb/9780198832232.013.30

Kügler, F. & Féry, C. (2017). Postfocal downstep in German. *Language and Speech, 60*(2), 260–288. https://doi.org/10.1177/0023830916647204

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., Jannedy, S., Michalsky, J., Niebuhr, O. & Peters, J. (2015). DIMA: Annotation guidelines for German intonation. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. IPA Public Archive.

Kushch, O., Igualada, A. & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience, 33*(8), 992-1004. https://doi.org/10.1080/23273798.2018.1435894

Ladd, D. R. (1980). *The structure of intonational meaning: Evidence from English*. Indiana University Press.

Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

Ladd, D.R. & Morton, R. (1997). The Perception of Intonational Emphasis: Continuous or Categorical? *Journal of Phonetics, 25*, 313-342. https://doi.org/10.1006/jpho.1997.0046

Ladewig, S. H., & Bressem, J. (2013). A linguistic perspective on the notation of gesture phases. . In C. Müller, E. Fricke, A. Cienki, D. McNeill, S. Ladewig, & S. Tessendorf (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK), 38/1* (pp. 1060–1079). http://doi.org/10.1515/9783110261318.1060

Lambrecht, K., (1994). *Information structure and sentence form: Topic, focus and the mental representations of discourse referents*. Cambridge University Press.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior research methods, 41*, 841–849. https://doi.org/10.3758/BRM.41.3.841

Lenth, R. V. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.7.4-.1. <https://CRAN.R-project.org/package=emmeans>.

Leonard, T. & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes, 26*(10), 1457–1471. https://doi.org/10.1080/01690965.2010.500218

Levinson, S.C. & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers of Psychology, 6*, 731. http://doi.org/10.3389/fpsyg.2015.00731

Levy, E. T., & Fowler, C. A. (2000). The role of gestures and other graded language forms in the grounding of reference. In D. McNeill (Ed.), *Language and Gesture* (pp. 215-234). Cambridge University Press. http://doi.org/10.1017/cbo9780511620850.014

Llanes-Coromina, J., Prieto, P. & Rohrer, P. L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński & D. Śledziński (Eds.), *Speech Prosody 2018* (pp. 498-502). ISCA Archive.

Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J. & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology, 172*, 168–188. https://doi.org/10.1016/j.jecp.2018.02.004

Li, P., Baills, F. & Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition, 42*(5), 1015–1039. http://dx.doi.org/10.1017/S0272263120000054

Liberman, M. Y. (1975). *The intonational system of English* [Unpublished Doctoral Dissertation]. Massachusetts Institute of Technology..

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8*(2), 249–336.

Loehr, D. P. (2004). *Gesture and intonation* [Unpublished Doctoral Dissertation]. Georgetown University.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture, 7*(2), 179–214. https://doi.org/10.1075/gest.7.2.04loe

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology, 3*(1), 71–89. https://doi.org/10.1515/lp-2012-0006

Lopez-Ozieblo, R. (2020). Proposing a revised functional classification of pragmatic gestures. *Lingua, 247*, 102870. https://doi.org/10.1016/j.lingua.2020.102870

Lücking, A., Bergman, K., Hahn, F., Kopp, S. & Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces, 7*(1-2), 5–18. https://doi.org/10.1007/s12193-012-0106-8

Macedonia, M. (2014). Bringing back the body into the mind: gestures enhance word learning in foreign language. *Frontiers*

in Psychology, 5, 1467. https://doi.org/10.3389/fpsyg.2014.01467

Macoun, A. & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development, 40,* 68–81. https://doi.org/10.1016/j.cogdev.2016.08.005

Marslen-Wilson, W. D., Levy, E., & Komisarjevsky Tyler, L. (1982). Producing interpretable discourse: the establishment and maintenance of reference. In R. J. Jarvella and W. Klein (eds.), *Language, place, and action: Studies in deixis and related topics* (pp. 339-378). Wiley.

Mattiello, E. (2017). The popularisation of science via TED Talks. *International Journal of Language Studies, 11*(4), 77–106.

Mattiello, E. (2019). A corpus-based analysis of scientific TED Talks: Explaining cancer-related topics to non-experts. *Discourse, Context & Media, 28,* 60–68. https://doi.org/10.1016/j.dcm.2018.09.004

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). *Montreal Forced Aligner* [Computer program]. Version 1.0.

McClave, E. (1994). Gestural beats: The rhythm hypothesis. *Journal of Psycholinguistic Research, 23*(1), 45–66. https://doi.org/10.1007/BF02143175

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research*, *27*(1), 69–89. https://doi.org/10.1023/A:1023274823974

McNeil, N. M., Alibali, M. W. & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior, 24*(2), 131–150. https://doi.org/10.1023/A:1006657929803

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago press.

McNeill, D. (2000). Introduction. In D. McNeill (Ed.*), Language and Gesture* (pp. 1–10). Cambridge University Press. https://doi.org/10.1017/CBO9780511620850

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press. http://doi.org/10.7208/chicago/9780226514642.001.0001

McNeill, D. (2006). Gesture: a psycholinguistic approach. In K. Brown (Ed.), *The encyclopedia of language and linguistics* (pp. 58–66). Elsevier.

McNeill, D. & Levy, E. T. (1993). Cohesion and gesture. Discourse Processes, 16(4), 363–386. https://doi.org/10.1080/01638539309544845

Molnár, V. (2002). Contrast from a contrastive perspective. *Language and Computers, 39*, 147-161. https://doi.org/10.1163/9789004334250_010

Mondada, L. (2016). Challenges of multimodality: Language and the body in social interaction. *Journal of Sociolinguistics, 20*(3), 336–366. https://doi.org/10.1111/josl.1_12177

Morett, L. M. (2014). When hands speak louder than words: The role of gesture in the communication, encoding, and recall of words in a novel second language. *Modern Language Journal*, *98*(3), 834–853. https://doi.org/10.1111/modl.12125

Mücke, D. & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation-Is it mediated by accentuation? *Journal of Phonetics, 44*, 47–61. https://doi.org/10.1016/j.wocn.2014.02.003

Müller, C. (2004). Forms and uses of the Palm Up Open Hand: a case of a gesture family? In C. Müller & R. Posner (Eds.), *The semantics and pragmatics of everyday gestures* (pp. 233-256). Weidler.

Müller, C. (2018). Gesture and sign: Cataclysmic break or dynamic relations? *Frontiers in Psychology, 9*, 1651. https://doi.org/10.3389/fpsyg.2018.01651

Nakatsukasa, K. (2016). Efficacy of recasts and gestures on the acquisition of locative prepositions. *Studies in Second Language Acquisition, 38*(4), 771–799. https://www.jstor.org/stable/26330942

Nobe, S. (1996). *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production* [Unpublished Doctoral Dissertation]. University of Chicago.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction: using the hands to learn math. *Psychological Science, 25*(4), 903–910. https://doi.org/10.1177%2F0956797613518351

O'Dell, M. & Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *14th International Congress of Phonetic Sciences* (pp- 1075–1078). ICPhS Archive.

Özçaliskan, S. & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition, 96*(3), B101–B113. https://doi.org/10.1016/j.cognition.2005.01.001

Özyürek, A., Willems, R. M., Kita, S. & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience, 19*(4), 605–616. https://doi.org/10.1162/jocn.2007.19.4.605

Pasdeloup, V. (1990). *Modèle de règles rythmiques du français appliquées à la synthèse de la parole* [Unpublished Doctoral Dissertation]. Université d'Aix en Provence.

Pasdeloup, V. (1992). A prosodic model for French text-to-speech synthesis: A psycholinguistic approach. In G. Bailly, C. Benoît & T. R. Sawallis (Eds.), *Talking machines: Theories, models and designs* (pp. 335-349). Elsevier Science Publisher.

Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In N. Calzolari, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (pp. 831-836). European Language Resources Association (ELRA).

Perniss, P. (2018). Why we should study multimodal language. *Frontiers in Psychology, 9*, 1109. https://doi.org/10.3389/fpsyg.2018.01109

Pierrehumbert, J. & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan & M.E. Pollack (Ed.), *Intentions in communication* (pp. 271 – 311). The MIT Press.

Portes, C., D'Imperio, M. & Lancia, L. (2012). Positional constraints on the initial rise in French. In Q. Ma, H. Ding, & D. Hirst (Eds.), *Speech Prosody 2012* (pp. 563-566). ISCA Archive.

Post, B. (2000). *Tonal and phrasal structures in French intonation*. Thesus The Hague.

Pouw, W. & Dixon, J. A. (2019a). Entrainment and Modulation of Gesture-Speech Synchrony Under Delayed Auditory Feedback. *Cognitive Science, 43*(3), e12721. https://doi.org/10.1111/cogs.12721

Pouw, W. & Dixon, J. A. (2019b). Quantifying gesture-speech synchrony. In K. Rohlfing, A. Grimminger, & U. Mertens (Eds.), *Proceedings of the 6th Gesture and Speech in Interaction (GESPIN) Conference* (pp. 75–80). http://www.doi.org/10.17619/UNIPB/1-815

Pouw, W., Jaramillo, J. J., Ozyurek, A. & Dixon, J. A. (2020). Quasi-rhythmic features of hand gestures show unique modulations within languages: Evidence from bilingual speakers. *Proceedings of the 7th Gesture and Speech in Interaction (GESPIN) Conference*. http://www.wimpouw.com/PouwJaramilloOzyurekDixon_PP.pdf

Pouw, W., Proksch, S., Drijvers, L., Gamba, M., Holler, J., Kello, C., Schaefer, R. S. & Wiggins, G. A. (2021). Multilevel rhythms in multimodal communication. *Philosophical Transactions of the Royal Society B, 376*(1835), 20200334. https://doi.org/10.1098/rstb.2020.0334

Prieto, P., Cravotta, A., Kushch, O., Rohrer, P., and Vilà-Giménez, I. (2018). Deconstructing beat gestures: a labelling proposal. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński & D. Śledziński (Eds.), *Speech Prosody 2018*, (pp. 201–205). ISCA Archive.

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E. & Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics*, *49*, 41–54. https://doi.org/10.1016/j.wocn.2014.10.005

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 223-256). Academic Press.

Prince, E. F. (1992). The ZPG letter: subjects, definiteness and information status. In S. Thompson and W. Mann (Eds.), *Discourse description: Diverse analyses of a fund raising text* (pp. 295-325). John Benjamins.

Pronina, M., Hübscher, I., Holler, J., & Prieto, P. (2021). Interactional training interventions boost children's expressive pragmatic abilities: evidence from a novel multidimensional testing approach. *Cognitive Development, 57*, 101003. https://doi.org/10.1016/j.cogdev.2020.101003.

R Core Team (2021). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

Repp, S. (2010). Defining “contrast” as an information-structural notion in grammar. *Lingua, 120*, 1333–1345. https://doi.org/10.1016/j.lingua.2009.04.006

Riester, A. & Baumann, S. (2013). Focus Triggers and Focus Types from a Corpus Perspective. *Dialogue Discourse, 4*(2), 215–248. https://doi.org/10.5087/dad.2013.210

Riester, A., Lorenz, D. & Seemann, N. (2010). A recursive annotation scheme for referential information status. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)* (pp. 717-722). European Language Resources Association (ELRA).

Riester, A. & Piontek, J. (2015). Anarchy in the NP. When new nouns get deaccented and given nouns don’t. *Lingua, 165*, 230–253. https://doi.org/10.1016/j.lingua.2015.03.006

Ritz, J., Dipper, S. & Götze, M. (2008). Annotation of Information Structure: an Evaluation across different Types of Texts. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, and D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and*

Evaluation (LREC'08) (pp. 2137–2142). European Language Resources Association (ELRA).

Rochet-Capellan, A., Laboissière, R., Galván, A. & Schwartz, J.-L. (2008). The speech focus position effect on jaw-finger coordination in a pointing task. *Journal of Speech, Language, and Hearing Research , 51*(6), 1507–1521. https://doi.org/10.1044/1092-4388(2008/07-0173)

Rohlfing, K. J., Grimminger, A., & Lüke, C. (2017). An interactive view on the development of deictic pointing in infancy. *Frontiers in Psychology, 8*, 1319. https://doi.org/10.3389/fpsyg.2017.01319

Rohrer, P. L., Prieto, P. & Delais-Roussarie, E. (2019). Beat gestures and prosodic domain marking in French. In S. Calhoun, P. Escudero, M. Tabain & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1500-1504). Australasian Speech Science and Technology Association Inc.

Rohrer, P. L., Vilà-Giménez, I., Florit-Pons, J., Gurrado, G., Esteve-Gibert, N., Ren, A., Shattuck-Hufnagel, S., & Prieto, P. (2021, February 24). The MultiModal MultiDimensional (M3D) labeling system. https://doi.org/10.17605/OSF.IO/ANKDX

Roustan, B. & Dohen, M. (2010). Co-production of contrastive prosodic focus and manual gestures: Temporal coordination

and effects on the acoustic and articulatory correlates of focus. *Speech Prosody 2010*, 100110, 1–4.

Rusiewicz, H. L. (2010). *The role of prosodic stress and speech perturbation on the temporal synchronization of speech and deictic gestures* [Unpublished Doctoral Dissertation]. University of Pittsburgh.

Rusiewicz, H. L., Shaiman, S., Iverson, J. M. & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication, 57*, 283–300. https://doi.org/10.1016/j.specom.2013.06.004

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50*, 696–735. http://doi.org/10.1353/lan.1974.0010

Sandler, W. (2018). The body as evidence for the nature of language. *Frontiers in Psychology, 9*, 1782. http://doi.org/10.3389/fpsyg.2018.01782

Selkirk, E. O. (1978). On prosodic structure and its relation to syntactic structure. In T. Fretheim (Ed.), *Nordic prosody II* (pp. 111-140), TAPIR.

Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. The MIT Press.

Selkirk, E. (1995). Sentence Prosody: Intonation, Stress, and Phrasing. In J. A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 550–569). Blackwell.

Shattuck-Hufnagel, S., Ostendorf, M. & Ross, K. (1994). Stress shift and early pitch accent placement in lexical items in American English. *Journal of Phonetics, 22*(4), 357–388. https://doi.org/10.1016/S0095-4470(19)30291-8

Shattuck-Hufnagel, S., & Prieto, P. (2019). Dimensionalizing co-speech gestures. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1490–1494). Australasian Speech Science and Technology Association Inc.

Shattuck-Hufnagel, S. & Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.01514

Shattuck-Hufnagel, S., Ren, A., Mathew, M., Yuen, I. & Demuth, K. (2016). Non-referential gestures in adult and child speech: Are they prosodic? In J. Barnes, A. Brugos, S. Shattuck-Hufnagel & N. Veilleux (Eds.), *Speech Prosody 2016* (pp. 836–839). ISCA Archive. https://doi.org/10.21437/SpeechProsody.2016-171

Shattuck-Hufnagel, S., Ren, P. L., & Tauscher, E. (2010). Are torso movements during speech timed with intonational phrases?

Proceedings of the International Conference on Speech Prosody, 1–4. ISCA Archive.

Shattuck-Hufnagel, S. & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247. https://doi.org/10.1007/BF01708572

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. *Second International Conference on Spoken Language Processing*. European Language Resources Association (ELRA).

Silverman, K. E. & Pierrehumbert, J. B. (1990). The timing of prenuclear high accents in English. *Papers in laboratory phonology I*, 72–106. https://doi.org/10.1017/CBO9780511627736.005

Skipper, J. I. (2014). Echoes of the spoken past: how auditory cortex hears context during speech perception. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1651), 20130297. https://doi.org/10.1098/rstb.2013.0297

Skopeteas, S., Fiedler, I., Hellmuth, S., Schwarz, A., Stoel, R., Fanselow, G., Féry, C. & Krifka, M. (2006). *Questionnaire on information structure (QUIS): reference manual* (Vol. 4). Universitätsverlag Potsdam.

Steedman, M. (2014). The surface-compositional semantics of English intonation. *Language*, *90*(1), 2–57. https://doi.org/10.1353/lan.2014.0010

Stewart, M. G. (2010, February). *How YouTube thinks about copyright.* [Video]. TED Conferences. https://www.ted.com/talks/margaret_gould_stewart_how_youtube_thinks_about_copyright

Swerts, M., & Krahmer, E. (2010). Visual prosody of newsreaders: Effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics, 38*, 197–206. https://doi.org/10.1016/j.wocn.2009.10.002

TEDx Talks (2015, April 14). *Comment oser prendre des risques pour vivre une vie intense ?* | Frederique Bedos [video]. YouTube. https://www.youtube.com/watch?v=zjEfmB0DV0k

TEDx Talks (2016, July 5). *Et si votre rêve devenait possible...* | *David Laroche* [video]. YouTube. https://www.youtube.com/watch?v=fbGfe78_2jc

TEDx Talks. (2018, December 18). *Négociation: ne cherchez pas le compromis* | *Julien Pelabere* [Video]. YouTube. https://www.youtube.com/watch?v=N9duDfWSfU4

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture, 8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel

Turk, O. (2020). *Gesture, prosody and information structure synchronisation in Turkish* [Unpublished Doctoral Dissertation]. Victoria University of Wellington.

Veilleux, N., Shattuck-Hufnagel, S. & Brugos, A. (2006). Transcribing Prosodic Structure of Spoken Utterances with ToBI. MITOpenCourseware. Retrieved from https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/index.htm

Vilà-Giménez, I., Dowling, N., Demir-Lira, Ö. E., Prieto, P., & Goldin-Meadow, S. (2021). The predictive value of non-referential beat gestures: Early use in parent-child interactions predicts narrative abilities at 5 years of age. *Child Development*, *92*(6), 2335–2355. http://doi.org/10.1111/cdev.13583

Vilà-Giménez, I., Florit-Pons, J., Rohrer, P. L., Muñoz-Coego, S., Gurrado, G., & Prieto, P. (2022, June 17). Audiovisual corpus of Catalan children's narrative discourse development. https://doi.org/10.17605/OSF.IO/NPZ3W

Vilà-Giménez, I. & Prieto, P. (2021). The value of non-referential gestures: a systematic review of their cognitive and linguistic

effects in children's language development. *Children, 8,* 148. https://doi.org/10.3390/children8020148

Voeten, C. (2022). *buildmer: Stepwise elimination and Term Reordering for Mixed-Effects Regression*. R package version 2.4, <https://CRAN.R-project.org/package=buildmer>.

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wagner, S. M., Nusbaum, H. & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language, 50*(4), 395–407. https://doi.org/10.1016/j.jml.2004.01.002

Weisberg, J., Hubbard, A. L. & Emmorey, K. (2017). Multimodal integration of spontaneously produced representational co-speech gestures: an fMRI study. *Language, Cognition and Neuroscience*, *32*(2), 158–174. https://doi.org/10.1080/23273798.2016.1245426

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Wolf, D., Rekittke, L.-M., Mittelberg, I., Klasen, M., & Mathiak, D. (2017). Perceived conventionality in co-speech gestures involves the fronto-temporal language network. *Frontiers in Human Neuroscience, 11*, 573. https://doi.org/10.3389/fnhum.2017.00573

Xu, Y. & Xu, C. X. (2005). Phonetic realization of focus in English declarative intonation. *Journal of Phonetics, 33*(2), 159–197. https://doi.org/10.1016/j.wocn.2004.11.001

Yap, D. & Casasanto, D. (2018). Beat gestures encode spatial semantics. In C. Kalish, M. A. Rau, X. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1211). Cognitive Science Society.

Yasinnik, Y., Renwick, M. & Shattuck-Hufnagel, S. (2004). The timing of speech-accompanying gestures with respect to prosody. *Proceedings of the International Conference: From Sound to Sense, 50*, 10–15.

Yoshioka, K. (2008). Gesture and information structure in first and second language. *Gesture, 8*, 236–255. http://doi.org/10.1075/gest.8.2.07yos

DOCTORAT EDUCATION
BRETAGNE LANGAGES, INTERACTIONS
LOIRE COGNITION, CLINIQUE

Nantes Université

Universitat Pompeu Fabra Barcelona

Titre : Une analyse temporelle et pragmatique de l'association geste-parole : Une approche basée sur un corpus utilisant le nouveau système d'annotation MultiModal MultiDimensional (M3D)

Mots clés : Geste, Prosodie, Structure Informationnelle, Synchronie Geste-Parole

Résumé : Le langage est essentiellement multimodal. En effet, des études récentes ont montré à la fois la forte relation temporelle entre les gestes co-verbaux et la proéminence prosodique et leur pertinence pragmatique. Cependant, ces études ont eu tendance à se concentrer sur le rôle de la proéminence prosodique en tant qu'attracteur principal pour la production de gestes, et peu de recherches empiriques ont évalué d'une manière systématique le rôle de la structure phrastique prosodique, ou la contribution conjointe de la proéminence gestuelle et prosodique pour des effets pragmatiques, en particulier en tant comme marqueurs de la structure informationnelle. En outre, aucune étude n'a pris en compte la différence potentielle entre les gestes référentiels et les gestes non référentiels.

Une analyse multidimensionnelle des différents traits du geste est cruciale pour permettre une évaluation systématique de leurs caractéristiques prosodiques et pragmatiques. Cette thèse poursuit un double objectif. Tout d'abord, elle propose une nouvelle approche de l'annotation des gestes co-verbaux qui épouse une vision dimensionnelle, selon laquelle les chercheurs devraient considérer les caractéristiques sémantiques, pragmatiques et prosodiques des gestes d'une manière non mutuellement exclusive. En second lieu, cette thèse vise à mieux comprendre les relations prosodiques et pragmatiques des gestes référentiels et non référentiels, en particulier la façon dont la structure phrastique prosodique influence les modèles de production gestuelle, et comment ces deux modes de communication interagissent pour des raisons pragmatiques.

Title : A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system

Keywords : Gesture, Prosody, Information Structure, Gesture-Speech Synchrony

Short Abstract : Human language is essentially multimodal and recent studies within the field of gesture research have shown both the strong temporal relationship between manual co-speech gestures and prosodic prominence, and have given initial evidence of the relevant pragmatic role of gestures. However, studies have tended to focus on the role of prosodic prominence alone as the main attractor for gesture production, and little empirical research has systematically assessed the role of prosodic phrasal structure in the attraction of gesture, or the joint contribution of gestural and prosodic prominence for pragmatic effects, particularly in terms of signaling information structure. Furthermore, no studies have specifically accounted for potential difference between referential and non-referential gestures.

A multidimensional analysis of independent aspects of gesture is crucial to allow for a systematic assessment of their different prosodic and pragmatic characteristics. The thesis contains two main objectives. First, it proposes a novel gesture labeling system (i.e., the MultiModal MultiDimensional (M3D) system) according to which the semantic, pragmatic, and prosodic characteristics of gestures should be assessed in a non-mutually exclusive manner. Second, this thesis applies the system to better understand the prosodic and pragmatic characteristics of both referential and non-referential gestures, particularly in terms of how phrasal prosodic structure influences gestural production patterns, and how these two modes of communication interact for pragmatic effect.