Temporal structure in language production and processing: a crossmodal comparison of spoken and sign language

Doctoral dissertation by:

Chiara Luna Rivolta

Supervised by:

Dr. Manuel Carreiras and Dr. Brendan Costello

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

2023

The project that gave rise to these results received support of a fellowship from "La Caixa" Foundation and the European Union's Horizon 2020 research and innovation programme under the Marie-Skłodowska-Curie grant agreement no. 713673 (personal fellowship code LCF/BQ/DI18/11660008).

bcbl
BASQUE CENTER
ON COGNITION, BRAIN
AND LANGUAGE

Paseo Mikeletegi, 69, Donostia-San Sebastián
January, 2023

Aknowledgements

This PhD has been a four years long adventure and now that it is coming to an end, when I look back, what I remember most fondly are all the people who have been a part of it.

I don't think I have enough *thank you* for all of you, honestly.

I would like to start by thanking the colleagues and fellow researchers who made this thesis possible.

My first thank you goes to my supervisors. To Dr. Manuel Carreiras, thank you for your guidance and your comments (event the harsh ones!) that pushed me to do better and better. To Dr. Brendan Costello, I feel so lucky to have had you as my supervisor for the last four years. Thank you for teaching me about sign language, for giving me such a good and healthy example of mentorship, and for letting me grow into the researcher I am today. You are an incredible model, both personally and academically.

I also want to thank Dr. Mikel Lizarazu, for his patience in teaching me all the wonders of MEG. Romain and Martin, the Kinect team, for the long days spent trying to make our motion tracking lab work.

My gratitude also goes to Dr. Asli Ozyurek, Dr. Wim Pouw and Dr. Linda Drijvers, for welcoming me in their team and giving me the chance to learn so much from them.

To the Neurolangers, thank you for the stimulating conversations during our meetings. A global thank you goes to all the BCBLians, for providing such a good working environment. A special thanks to Eider and Ana for the constant support with all the head-breaking bureaucracy and the amazing lab team (especially Araitz, Manex, Leire, Maite, David and Ainhoa) for making the long testing hours more fun. Finally I want to thank the predoc community for the support and the friendship you showed me.

I am also very thankful to all my friends that accompanied me, outside BCBL (and sometimes both).

To Alberto goes my deepest love, we started this adventure together and I can’t imagine doing this PhD without you by my side. Thank you for being such a good friend, an incredible advisor and a role model for me.

A Asi, gracias por estar siempre en mi equipo, pase lo que pase. La verdad es que no puedo imaginar mi vida en Donosti sin ti en ella.

To Irene, Eneko and Dani (and Biruji, of course!), thank you for your unconditioned friendship and for teaching me something new every single day.

To my *koadrilla donostiarra*, so big and beautiful. When moving here four years ago I never imagined that I was going to find such an amazing group of friends. Thank you for all the laughs, the extremely long coffee at Bizi, the holidays and the uncountable drinks and dinners. Thank you for being there in the good moments, and in the not-so-good ones. I want you to know that each one of you left a mark. Thank you Christoforos, Vicente, Hana, Catherine, Abraham, Giorgio, Laura, Jordi, Inés, Marta, Jose P, Pierma, Carlos, Stefano, Candice, Conrad, Jessi, Trisha, Jose A, Polina, Giulia, Sandy and Dani.

A mis chicas, Raquel Elena y Eleonora, gracias por haber sido mi casa (literalmente) en Donosti. No hubiera podido pedir unas compañeras de piso y de cuarentena mejores de vosotras.

A mi talde de AEK, y especialmente a Maialen, para haberme enseñado lo bonito y divertito que es aprender Euskera. Milesker.

To Jacopo, il mio compagno di avventure, and all my master friends. I fell lucky to have started my journey into research with you.

I want thank you my Italian families too. Mamma, papá e Fede: grazie per avermi dato la libertá di viaggare e vivere in un paese straniero, facendomi sentire amata anche da lontano.

Alla mia seconda famiglia, quella che scegli e che ti sceglie: Anna e Cristiana, Giacomo, Paolo, Dario, and Gabriele. Non ve lo dico abbastanza, ma a volte é davvero difficile diventare grande lontano da voi. Grazie perché non abbiamo bisogno di ripeterlo spesso per essere sicuri che ci saremo sempre. Vi voglio bene.

A Silvia, che probabilmente non lo sa però, davvero, non ce l’avrei fatta ad arrivare alla linea del traguardo senza il suo aiuto.

Resumen

Capítulo 1

En nuestro día a día percibimos todo, sea un evento interno o externo, a través de un filtro temporal, incluidos los estímulos muy complejos, como el caso del lenguaje. Una parte significativa de la literatura sobre las lenguas habladas se dedica a la estructura temporal del lenguaje, y el estudio del habla demuestra —a través de medidas conductuales y neurofisiológicas — la periodicidad en su estructura temporal. Esta regularidad es evidente tanto en la percepción como en la producción del lenguaje: la medición de los movimientos de los articuladores del habla y los análisis del envolvente del habla han revelado una regularidad temporal en el rango de frecuencia de 4-5 Hz (Drullman, 2019; Goswami & Leong, 2013; Walsh & Smith, 2002). También se encuentran regularidades similares a nivel cerebral: durante la percepción del habla, las poblaciones neuronales oscilan en frecuencias situadas dentro de las bandas delta (< 4 Hz), asociadas con el contorno prosódico, y theta (4-8 Hz), que corresponden a la frecuencia de las sílabas (Meyer, 2018). Algunos investigadores sostienen la idea de que la periodicidad en el habla temporal favorece la descodificación eficaz de la información lingüística a partir de la señal acústica, por lo que tendría un papel fundamental en el procesamiento del lenguaje (Doelling et al., 2014; Thomson & Goswami, 2008).

Cuando pensamos en el lenguaje, a menudo lo asociamos con el habla; no obstante, el lenguaje se puede expresar a través de otros canales además del acústico. Las personas sordas se comunican mediante la lengua de signos, que emplea únicamente el canal visual. Tanto la lengua hablada como la lengua de signos presentan la misma complejidad en cuanto a su análisis lingüístico: es decir, a nivel fonológico, morfológico, sintáctico y semántico (MacSweeney et al., 2008; Pfau et al., 2012; Sandler & Lillo-Martin, 2006). En la lengua de signos la información se transmite mediante el uso de diferentes articuladores, tanto manuales (manos y dedos) como no manuales (el torso, la cabeza, la mirada, las cejas y la boca). Dentro del componente puramente manual de la señal se identifican tres parámetros fonológicos principales cuya combinación produce la unidad léxica del signo: la configuración de la mano, la localización del signo y el movimiento que la mano realiza en el espacio (Brentari, 1998; Herrero Blanco, 2009). Los componentes no manuales también pueden aportar información (sub)léxica o de otro tipo: por ejemplo, las cejas transmiten información prosódica, mientras que los movimientos del torso

marcan la toma de turnos en un discurso.

Aunque las lenguas de signos y las lenguas habladas tengan muchas similitudes, su organización temporal es muy distinta y está condicionada por su modalidad; mientras que la modalidad acústica se caracteriza por su alta resolución temporal, la modalidad visual favorece la información espacial sobre la temporal (Holcombe, 2009; Meier, 2002), y estas propiedades se reflejan en la estructura de la señal lingüística. Una diferencia entre las lenguas de signos y las habladas reside en el ritmo de producción de las unidades lingüísticas. La duración de los signos suele ser el doble (~2 signos por segundo) de la duración de las palabras (~4-5 palabras por segundo), y este efecto se ha atribuido al mayor tamaño de los articuladores y los movimientos de la lengua de signos (Grosjean, 1977a; Klima, E. S., & Bellugi, 1979; Wilbur, 2009). Otra distinción importante es que la lengua de signos tiende a favorecer una presentación paralela y multidimensional de la información: los distintos articuladores se mueven simultáneamente, y los tres parámetros fonológicos se realizan a la vez para crear el signo. El lenguaje hablado, aunque incluya algunos flujos de información paralelos (por ejemplo, los gestos que acompañan al habla), se aproxima más a una interfaz serial y unimodal: las lenguas habladas se estructuran y ordenan de manera secuencial.

El objetivo de esta tesis es investigar y esclarecer el impacto de la modalidad (visual o acústica) en la organización temporal del lenguaje, y cómo esto afecta a la forma en que nuestro cerebro y nuestro sistema cognitivo procesan el lenguaje. La lengua oral, que utiliza el canal auditivo, y la lengua de signos, que se expresa únicamente mediante el canal visual, son dos ejemplos perfectos para aislar el efecto que la modalidad tiene sobre la lengua. En este trabajo doctoral comparo el español europeo y la lengua de signos española (LSE) en tres estudios distintos que se enfocan en diferentes niveles del lenguaje: la comprensión (capítulo 2), la producción (capítulo 3) y el procesamiento neural (capítulo 4).

Capítulo 2

El objetivo principal del primer estudio presentado en este trabajo doctoral es investigar cómo la estructura temporal de la lengua afecta a su comprensión. Para averiguar la contribución de, por un lado, la modalidad perceptiva y, por otra, de la estructura lingüística, hemos empleado tres tipos de estímulos: oraciones semánticamente impredecibles en LSE y en español y estímulos visuales no lingüísticos (que consistían en vídeos de un punto trazando símbolos en la pantalla).

Hemos adaptado un paradigma utilizado extensivamente en la literatura de la lengua hablada, locally time-reversed speech paradigm (Greenberg & Arai, 2001; Steffen, A., & Werani, 1994; Stilp, Kiefte, Alexander, & Kluender, 2010) a la modalidad visual. Este paradigma consiste en dividir la señal (acústica o visual) en segmentos, invirtiendo la señal de cada uno pero manteniendo el orden de los segmentos. La duración del segmento seleccionado afecta al nivel de distorsión: cuanto más grande el segmento, mayor el grado de distorsión. Hemos aplicado 6 niveles de distorsión (la señal original y 5 niveles de distorsión) para presentar los estímulos modificados a 25 bilingües bimodales, personas oyentes cuya lengua nativa es el español y que tienen un nivel muy alto de LSE. La tarea experimental consistía en reproducir el mayor número de palabras, signos o símbolos en el orden correcto.

Los resultados de este estudio muestran que la comprensión de la lengua oral es muy alta con niveles de distorsión con segmentos de hasta 40 ms, y a partir de ese umbral la comprensión baja muy rápidamente hasta perderse por completo. Estos resultados reproducen perfectamente los de otros estudios con otras lenguas orales (Ueda, Nakajima, Ellermeier, & Kattner, 2017). En las tareas visuales (LSE y estímulos no lingüísticos), la capacidad para entender y reproducir el material también disminuye con mayores niveles de distorsión, pero de una forma mucho más gradual; no existe un umbral que marque una clara caída en el nivel de comprensión. Además, en comparación con la condición no lingüística, la LSE es más resistente a la distorsión temporal: los signos se entienden a todos los niveles de distorsión y, aun en el nivel de distorsión máxima, se podía reconocer la mitad de los signos.

Estos resultados demuestran que la LSE es mucho más resistente a la distorsión temporal de la información lingüística que la lengua oral. La modalidad visual permite y fomenta la presentación paralela de la información lingüística: esto es evidente en el movimiento simultáneo de los articuladores y en algunos parámetros fonológicos que, por su propia naturaleza, son muy estáticos (por ejemplo, la configuración de la mano y su ubicación en el espacio). Esta presentación paralela crea redundancia (y, como consecuencia, una sobrerrepresentación) de la información lingüística.

Capítulo 3

Para estudiar las propiedades físicas de la señal visual del lenguaje (tanto signado como oral), hemos utilizado una herramienta de motion tracking o captura de movimiento llamada Kinect. Esta herramienta permite detectar y grabar los movimientos de distintas partes del cuerpo y de la cara en un espacio tridimensional. Hemos empleado esta herramienta para grabar vídeos de narración natural en español, ruso, LSE y lengua de signos rusa (RSL), para su posterior uso como estímulos para el estudio presentado en el capítulo 4. El objetivo de este estudio es valorar la idoneidad de diferentes medidas cinemáticas para describir las propiedades de la señal visual, y averiguar si estas medidas permiten distinguir entre los movimientos corporales de las lenguas orales y los de las lenguas de signos.

Las medidas que hemos empleado son: número de movimientos, ritmicidad, espacio de movimiento, magnitud de movimiento, análisis de tiempo-frecuencia y UMAP (una técnica empleada para el agrupamiento automático de datos). Después de limpiar y preparar los datos de motion tracking, hemos calculado estas medidas para los movimientos de cinco partes de cuerpo seleccionadas por su relevancia lingüística en lengua de signos (cabeza, torso, mano derecha, hombro derecho y mano izquierda) para examinar los resultados de las dos lenguas habladas y las dos lenguas de signos.

En cuanto a los resultados, la mayoría de estas medidas consiguen encontrar claras diferencias entre las dos lenguas. Los movimientos corporales de la lengua de signos se caracterizan por una mayor periodicidad, un mayor uso del espacio y un mayor nivel de homogeneidad, en comparación con los movimientos de la lengua hablada. Además, los articuladores de la lengua de signos no muestran el mismo patrón cinemático entre sí, lo que destaca la importancia de seguir investigando para desentrañar la contribución específica de cada articulador.

Aun basándose en el movimiento corporal y, por ende, en la misma modalidad (visual), la señal visual de la lengua de signos y la de la lengua oral tienen estructuras temporales distintas. En las lenguas orales, los movimientos del cuerpo se producen simultáneamente con el habla. Estas dos señales, que pertenecen a dos modalidades (la visual y la acústica) caracterizadas por estructuras temporales muy diferentes, necesitan interactuar y combinarse para crear un mensaje lingüístico cohesivo. El habla tiene un rol más importante en la comunicación y, como consecuencia, la señal visual se adapta a los patrones temporales de esta (Wagner, Malisz, & Kopp, 2014). En cambio, en la lengua de signos, la señal es transmitida únicamente a través del canal visual y es interesante observar que, en ausencia de la modalidad acústica, los movimientos corporales tienden a organizarse temporalmente de manera diferente. El análisis exploratorio presentado en este capítulo evidencia la necesidad de una descripción sistemática de las propiedades físicas de la señal de la lengua de signos y las diferencias entre sus articuladores.

Capítulo 4

El fenómeno denominado entrainment se refiere a la sincronización entre las oscilaciones propias del habla y la actividad de las neuronas durante el procesamiento del lenguaje (Obleser & Kayser, 2019). Este fenómeno se estudia ampliamente en el caso de las lenguas orales; con este estudio pretendemos investigar si este fenómeno ocurre solo en la lengua hablada, o si se extiende a una lengua expresada y percibida a través de la modalidad visual, como la lengua de signos. Esta sincronización ocurre de forma automática con cualquier tipo de señal casi periódica (incluidas las lenguas desconocidas), pero está condicionada por procesos descendentes (top-down) como la atención o el conocimiento de la lengua. Todavía se desconoce la naturaleza del impacto de estos factores: algunos estudios han mostrado diferencias en la fuerza de la sincronización, otros en las frecuencias que caracterizan la sincronización y otros en la distribución topográfica en el cerebro (Brookshire et al., 2017; Ding et al., 2016; Lizarazu et al., 2021; Peña & Melloni, 2012).

Hemos utilizado la magnetoencefalografía (MEG) para registrar la actividad neurofisiológica de dos grupos de participantes oyentes —signantes expertos y personas sin conocimientos de la lengua de signos— mientras veían vídeos en una de cuatro lenguas: una lengua hablada conocida (español), una lengua hablada desconocida (ruso), una lengua de signos conocida para las personas signantes (LSE) y una lengua de signos desconocida (RSL). Utilizamos una tarea ortogonal para asegurarnos de que los participantes prestasen atención a los vídeos. Este diseño nos ha permitido investigar el efecto específico de la modalidad lingüística y del conocimiento previo de la lengua sobre el entrainment.

Los resultados de este experimento muestran que la actividad cerebral se sincroniza tanto con el envolvente del habla como con los movimientos de la mano derecha en la lengua de signos. Para las lenguas habladas replicamos los resultados típicos encontrados en investigaciones anteriores: sincronización en las bandas de frecuencia delta y theta localizada en regiones temporales bilaterales (Bourguignon et al., 2013; Ding, Melloni, et al., 2017; Keitel et al., 2017). El entrainment con la lengua de signos depende de las propiedades específicas de su estructura temporal: está restringido a las frecuencias bajas (de la banda delta), en consonancia con la periodicidad más lenta de la señal visual, y se localiza en las áreas parietales derechas (asociadas con el procesamiento del movimiento). La sincronización es mucho más fuerte para las lenguas orales en comparación con las lenguas de signos. El conocimiento previo de la lengua afecta a la sincronización del cerebro al input lingüístico: en la lengua de signos, la lengua conocida produce más sincronización que la lengua desconocida, pero el patrón se invierte en las lenguas orales. Por último, los resultados sugieren que la experiencia con una lengua de signos afecta a la forma de procesar el habla en cuanto a la sincronización con la señal audiovisual hablada.

Capítulo 5

Gracias a este trabajo doctoral hemos averiguado que la modalidad (acústica y visual) afecta no solo a la forma en que producimos el lenguaje (y, en particular, a su estructura temporal), sino también cómo nuestro cerebro emplea distintos mecanismos para percibirla y procesarla. El ámbito visual favorece la presentación de fuentes de información simultáneas, pero espacialmente distintas. En la lengua de signos, esto queda evidenciado por el uso de varios articuladores distintos que se mueven de forma simultánea y casi independiente. La lengua de signos es una señal multicapa de la cual podemos aislar la información procedente de cada articulador (o capa), mientras que el habla se percibe como una señal compleja y unificada que se va modulando en la dimensión temporal.

El análisis de motion tracking presentado en el capítulo 3 demuestra la eficacia de esta técnica para captar la dinámica temporal de la señal visual de la lengua de signos. Además, esta dinámica es diferente en los cinco articuladores investigados; este resultado resalta la importancia de la naturaleza multicapa de la lengua de signos: cada articulador no solo transmite información diferente, sino que también exhibe diferentes propiedades cinemáticas en comparación con otras partes del cuerpo. Al comparar la señal visual de la lengua oral y la de la lengua de signos, encontramos que tienen características cinemáticas diferentes. En las lenguas habladas, los movimientos corporales que acompañan al habla se acomodan a las propiedades temporales de la modalidad acústica empleada por el habla, que representa la fuente primaria de información lingüística. Por el contrario, la lengua de signos emplea únicamente la modalidad visual y la señal se organiza según las propiedades del sistema visual.

Para ser comprendida, la señal lingüística debe ser procesada por nuestro sistema cognitivo. Nuestro cerebro dispone de diferentes sistemas perceptivos en función de la modalidad utilizada para recibir la información, y muestra una predisposición a explotar la periodicidad de los estímulos para un procesamiento eficiente de la información. Nuestros resultados dejan patente que la estructura temporal es mucho menos importante para la lengua de signos que para la lengua oral; esto es evidente tanto a nivel de comprensión, donde la lengua de signos es más resistente a la distorsión temporal de la señal (capítulo 2), como a nivel de procesamiento, donde el entrainment es menos fuerte que el observado en la lengua oral (capítulo 4).

Las diferencias entre la lengua de signos y la lengua oral que encontramos en la producción y la percepción del lenguaje están estrechamente relacionadas. Tanto las lenguas de signos como las lenguas orales han evolucionado de forma natural aprovechando las limitaciones y los puntos fuertes del sistema perceptivo que emplean. La señal acústica del habla es más rítmica, ya que la periodicidad en su estructura temporal es muy importante para su percepción y procesamiento. Para procesar la lengua de signos el sistema cognitivo utiliza la estructura temporal hasta cierto punto, pero de una forma cualitativamente y cuantitativamente diferente en comparación con la lengua oral. El procesamiento de la lengua de signos aprovecha otras características de la señal que no son captadas por su estructura temporal, como el dominio espacial. Sin embargo, nuestro cerebro procesa la información en el tiempo, es decir, a través de un filtro temporal. En consecuencia, la lengua de signos representa un caso especialmente interesante en el que la organización espacial convive e interactúa con la estructura temporal de la señal.

Abastract

The way we produce and perceive language is tightly related to the modality employed by the language itself. This doctoral thesis focuses on one specific language feature which is directly affected by perceptual modality: temporal structure. Spoken and signed languages, which use the acoustic and the visual modality, respectively, to express linguistic content, represent the perfect test case to investigate modality effects on the temporal organization of language. This thesis compares spoken European Spanish and Spanish Sign Language (LSE) in a bimodal bilingual population across different levels of processing: comprehension, production and neural processing.

To investigate how much language intelligibility relies on its temporal structure, we applied a temporal distortion using the locally time-reversed speech paradigm to spoken Spanish, LSE and a visual non-linguistic signal. We found that overall the visual modality is much more resilient to temporal distortion: Spanish has a distortion threshold after which intelligibility is almost completely lost whereas LSE is characterized by a gradual and constant decrease in intelligibility which never goes below 50%. The comparison between LSE and the non-linguistic visual stimuli reveals overall better performance in LSE, demonstrating that the presence of linguistic structure in the signal aids the decoding of distorted information. These results suggest that modality poses some constraints on the way temporally distorted information can still be retrieved and processed, and that the presence of linguistic structure aids the decoding of information.

To characterize the physical properties of the visual linguistic signal produced in spoken and sign language, we developed a custom-built motion tracking system that can record and measure the movement of different body and face parts over time. We used different kinematic variables adopted from the literature on gesture and speech analysis and assessed their suitability for describing the linguistic visual signal and for distinguishing between spoken and sign language. The analysis shows that the signed and spoken linguistic signals have different kinematic profiles: sign language is more rhythmic and employs faster movements which encompass a larger motion space. The analysis also reveals that the movement of different body parts, or articulators, is characterized by distinct temporal structures. These findings highlight the potential of motion tracking analysis to better understand sign language production and processing.

The periodicity in the spoken language signal has proven to be very important in language processing: brain waves synchronize with the frequencies presented in the speech envelope, aiding language comprehension. In the last study presented in this thesis we tested whether this phenomenon, known as language-brain entrainment, is employed in sign language processing and how much it is modulated by language knowledge. We recorded MEG activity of participants while they watched (and heard) naturalistic storytelling in four different languages: Spanish (known spoken language), Russian (unknown spoken language), LSE (known sign language) and Russian Sign Language (unknown sign language). We measured entrainment by calculating coherence between the brain activity recorded with MEG and the linguistic stimulus (characterized as the broadband speech amplitude envelope for spoken languages and as the right hand speed vector for signed languages). We reproduce the classical findings of entrainment in spoken language: synchronization in delta and theta frequency bands located in bilateral temporal regions. We find entrainment in sign language as well, but its characteristics are modulated by the properties of sign language: synchronization is limited to delta frequency band in motion processing areas and it is not as strong as in spoken language. The language knowledge modulates how the brain entrains to linguistic input but we found an interaction with modality, with opposite patterns in spoken and sign language.

This doctoral thesis focused on the effect of modality, acoustic and visual, on the temporal structure of language from two different but complementary approaches: language production and language perception. Taken together the results from the three studies converge in showing that modality shapes the characteristics of the temporal structure of language. In language production the physical properties of the linguistic signal organize according to the dominant modality used, as shown by the differences found between the sign language signal and the visual signal that accompanies speech. Overall, temporal structure plays a less important role in sign language compared to spoken language: the sign language signal is more resilient to temporal distortion and sign language processing does not rely on entrainment as much as spoken language does. The acoustic modality is specifically sensitive to the temporal resolution of the system; in contrast, the visual modality favors the spatial dimension to convey perceptual information. Language structure appears to optimally exploit the constraints and strengths of the perceptual system used to produce and perceive it.

Table of Contents

Chapter 1: Temporal structure of language

The perception of any type of event or information takes place over time, and our brain is predisposed to receive information organized at different time-scales and process it accordingly. This is especially the case during language processing. Our cognitive system needs to match with its own temporal mechanisms the complex temporal structure of the linguistic input, giving rise to effortless language comprehension. In this thesis I examine how the temporal properties of the linguistic signal influence the way we understand language and how the brain tunes to this temporal structure. The temporal organization of a signal is deeply connected with the sensory modality in which the signal is perceived; in this context, language offers a way to study this modulation through the comparison of languages which employ different modalities. In this work I will compare European Spanish, which is conveyed through the acoustic modality, and Spanish Sign Language (LSE), which uses the visual channel. This comparison allows us to further our understanding of the temporal structure of language and the way it adapts to our cognitive system.

In this first chapter of the thesis I introduce the overarching theoretical themes that guide this work. In section 1.1 I describe the fundamental role that temporal structure plays in language and provide an overview of the literature dedicated to this topic, mainly concentrated on spoken language. Section 1.2 offers the reader a general introduction to sign language; this section is in no way an exhaustive overview of the topic but focuses on those aspects of sign language which are relevant for this thesis. In section 1.3 I describe the various differences between temporal structure in sign and spoken language. Section 1.4 examines a specific phenomenon of interest in the study of language temporal structure, namely brain-language entrainment, which will be the focus of the main experiment of this thesis. Finally, in section 1.5 I lay out the main research questions that this work addresses as well as an overview of the experiments performed. The structure of the thesis is presented in section 1.6.

1.1 Why temporal structure is important for language processing

The human brain perceives and processes any incoming signal through the filter of temporal resolution, allowing it to integrate discrete units of signal into a continuous and unified percept. One clear example of this temporal filter at work in our everyday life is the spectrum of audible sounds. Dogs, for example, can hear sounds in frequencies up to 45kHz while for human beings the hearing range lies between 20 Hz and 20 kHz (and it slowly reduces with age). The temporal dimension is an integral part of what we perceive, while also representing the scaffolding structure of perception itself. The human brain does not have a single temporal resolution, but acts at multiple scales (Viemeister & Wakefield, 1991). On one hand, it is limited by the physiology of our nervous system, based on how quickly the system can encode the input (which, in turn, depends on specific properties of the nerves such as the duration of action potential spikes and subsequent refractory period). At the same time, it is also modulated by other factors such as the perceptual modality and the properties of the incoming information. Temporal resolution is usually higher in the auditory compared to the visual domain. The smallest detectable gaps in an auditory signal lie in the range of 2 to 20 ms (> 50 Hz), and this value depends on the spectral properties of the signal itself (Peters & Glasberq, 1993). A review of several studies on temporal resolution in the visual system identified two principal temporal scales (Holcombe, 2009): cognitive processes involving higher order visual perception (such as word recognition, motion and colour integration) are constrained at a slower temporal resolution (mostly below 4 Hz), while specialized and automatic mechanisms (such as flicker perception) work on a much faster scale of up to 50 Hz.

One type of complex signal which is subject to perception and processing is language. Research on language processing needs to take into account not only the filter of our cognitive system's temporal resolution, but the specific temporal structure of language itself, and how well they fit together. One way to examine this temporal structure is to investigate the effect of temporal distortion on language comprehension through the use of different temporal degradation paradigms, which tap into various properties of the speech signal. One example is a paradigm which consists of periodic interruption of speech by silence or modulated noise (Miller & Licklider, 1950). Several studies employing this type of distortion found a U-shaped intelligibility pattern: intelligibility is retained when the interruption/segmentation is applied at very high and very low rate, but in the range 2-4 Hz intelligibility is lost (Huggins, 1975; Jin & Nelson, 2010; Nelson & Jin, 2004; Saija et al., 2014; Shafiro et al., 2015). Shafiro and colleagues (2016) modulated the length of silent gaps by deleting the silence and concatenating the remaining speech segments or leaving the silent gap in the auditory signal. They found that performance was better when silence was maintained, and this was especially true for speech segmented in the critical range of 2-4 Hz. Another temporal distortion technique is to speed up the signal. Ghitza & Greenberg (2009) manipulated the speech rate with time-compression by a factor of three and inserted silent intervals of different lengths. Time-compression led to a loss of intelligibility and so performance falls at less than 50%. Interestingly, when 80 ms silent intervals were inserted periodically, such that the original temporal structure of the signal was restored, intelligibility was almost recovered. All these studies point to a specific temporal resolution window in speech that, when manipulated, impairs language comprehension.

Another widely used temporal distortion is locally time-reversed speech. In this paradigm the linguistic signal is first divided into segments or windows of a given duration and then each window is reversed while the global order of the windows is kept. This paradigm has been applied to several spoken languages, and typically the size of the window is manipulated to see how different window durations impact speech intelligibility (Ishida, 2021; Ishida et al., 2018; Kiss et al., 2008; Matsuo et al., 2020; Saberi & Perrott, 1999; Steffen, A., & Werani, 1994; Stilp et al., 2010; Teng et al., 2019; Ueda et al., 2017). All these studies found a very similar pattern of results: intelligibility is not affected by very small windows of less than 40 ms but starts to decrease when the reversal window is approximately 40 ms in size, is less than 50% when the window is between 60-70 ms, and for windows of 100 ms or longer speech comprehension is almost completely lost. Only two studies have applied the same paradigm to a signed language: Hwang (2011) for American Sign Language (ASL), and Rivolta et al., (2021) for LSE, which is reported in Chapter 2 of this thesis.

Speech is a quasi-periodic signal: it shows a certain level of regularity in its temporal structure which, even if not perfectly isochronous, can be assimilated to rhythm. This temporal regularity is evident in both behavioural and neurophysiological measures, and across perception and production domains. Measurement of speech articulator movements has revealed temporal regularity in the frequency range of 4-5 Hz (Lindblad et al., 1991; Riely & Smith, 2003; Walsh & Smith, 2002). Similarly, the acoustic analysis of the speech signal reveals a speech envelope with greater power for frequencies between 2 and 8 Hz and a peak between 4 and 5 Hz (Drullman, 2019; Goswami & Leong, 2013). Interestingly, similar regularities are found at the brain level as well: during speech perception neural populations oscillate at frequencies falling within the delta (< 4 Hz) and theta (4-8 Hz) bands, and, critically, are observed across different speakers, languages and speaking conditions (Meyer, 2018). A detailed explanation of this phenomenon, known as brain-language entrainment, and a review of the literature on this topic is provided in section 1.3 of this chapter. The consistency in this temporal periodicity across different languages and different linguistic domains points towards the existence of a common neural and cognitive mechanism devoted to processing language input as speech.

Although there is an extensive literature on the temporal structure of language (and a great degree of agreement on many of its properties), the specific function that temporal structure serves in language processing is still under debate. Several researchers draw a link between temporal regularities, both in terms of temporal window duration and dominant frequencies in the speech signal, and specific linguistic units. In particular, phonemes (Di Liberto et al., 2015; Lehongre et al., 2011), syllables (Doelling et al., 2014; Ghitza, 2013; Giraud & Poeppel, 2012; Greenberg & Arai, 2004; Howard & Poeppel, 2012; Luo et al., 2010) and prosodic phrases (Ding, Patel, et al., 2017; Molinaro & Lizarazu, 2018) have been proposed as linguistic correlates of periodicities in the temporal structure of speech. According to this view, the temporal regularity which characterizes the speech signal plays an extremely important role in language perception and comprehension: a quasi-periodic structure supports the efficient parsing and decoding of linguistic information from the acoustic signal. A failure to perceive and process this periodicity has been linked with reading deficits and poor language comprehension (Doelling et al., 2014; Thomson & Goswami, 2008). Alternatively, these temporal regularities found in language can be ascribed to the properties of the acoustic sensory channel (Cummins, 2012; Samuel, 1991). According to this view, periodicity does not reflect any linguistic segmentation of the signal: listeners simply pick up on temporally structured patterns in any type of acoustic signal in order to process it (Samuel, 2020). One way to examine this relationship between the temporal structure of the signal and its processing is to look at languages which do not employ the acoustic modality, such as signed languages.

1.2 Sign Language

When we think about language we often associate it with speech, and in general with spoken languages. These are languages that use primarily the acoustic modality both to produce and perceive the linguistic signal. Language is not limited to the oral channel though. When the acoustic modality is interrupted, for example in the deaf and hard-of-hearing population, language recruits the visual modality in the form of sign language. More than fifty years of extensive research from psycholinguists and linguists has shown that signed languages are fully-developed, natural languages with a syntactic and semantic structure comparable in complexity with that of spoken languages (MacSweeney et al., 2008; Pfau et al., 2012; Sandler & Lillo-Martin, 2006). Signed languages show the defining properties of language, such as duality of patterning, discreteness, and productivity, with the exception of the use of the vocal-auditory channel (Meier, 2002). Signed and spoken language also share the same acquisition process (Meier, 1991; Newport & Meier, 2018), including a specific critical period for acquisition (Mayberry & Fischer, 1989; Newport, 1990) and a manual babbling phase (Petitto & Marentette, 1991).

In spite of all the functional similarities between spoken and signed languages, they greatly differ on a formal and structural level due to the constraints posed by the use of different modalities (auditory-vocal and visual-gestural, respectively). In sign language the information is transmitted in the visual modality through the use of different articulators, both manual (hands and fingers) and non-manual (including the torso, head, eye-gaze, eyebrows and mouth). Within the purely manual component of the signal, phonological theories of sign language typically identify three main phonological parameters that make up the sign: the handshape (the form that the different fingers of the hand adopt), the location of the sign (on the signer's body or in the space in front of the body) and the movement performed by the hand(s) in the space (Brentari, 1998; Herrero Blanco, 2009; Sandler, 2011; Stokoe & Marschark, 2005). The combination of these parameters produces lexical signs. The non-manual components may also provide lexical information – a given sign may include a head tilt or a specific mouth pattern – but also operate at other levels of linguistic organization. Facial expressions and eyebrows movements carry the intonational component of prosody in sign languages (Reilly et al., 1990), while head movement can signal prosodic boundaries within a sentence (Nespor & Sandler, 1999). Torso displacement may serve to structure the discourse by marking role shift among different characters during indirect discourse (Cormier et al., 2013; Janzen, 2004; Lillo-Martin, 1995). These different sign language articulators all carry different types of linguistic information, and do so quasi-independently in a compositional manner.

In addition to the linguistic study of sign languages, which focuses on the properties and structure of the language itself, we can also ask how sign languages are processed and what cognitive processes are associated with this type of language. Various studies have investigated whether the psycholinguistic theories created for the spoken language domain apply to sign language as well. At the lexical level, in a recent paper Caselli and colleagues (2021) found evidence that sign recognition shares at least some of the same properties of word recognition: the spontaneous activation of phonological neighbours (i.e., similar lexical items) leads to competition among signs during lexical access. Moreover, sign neighbourhood density – the number of similar signs to the target sign – interacts with sign frequency as in spoken language. This finding is supported by a growing body of literature that shows phonological competition effects in sign language lexical access (Carreiras et al., 2008; Dye, 2006; Hildebrandt & Corina, 2002; Lieberman & Borovsky, 2020; Mayberry & Witcher, 2005; Meade et al., 2018; Villameriel et al., 2019). With regard to the neurobiology of language processing, a considerable body of research on the neural basis of sign language production and comprehension reveals substantial overlap in the language network areas for speech and sign (see Emmorey, 2021 for a review). Setting aside some differences in low-level processing, with signed languages activating occipital (visual) cortices and spoken languages activating superior temporal (auditory) cortical areas (Leonard et al., 2012), sign language processing mainly recruits the frontal and temporal regions of the left hemisphere associated with (spoken) language, as well as additional activation in the right hemisphere involved in the computation of spatial relationships.

The study of signed languages, alone and in comparison with spoken languages, aids our knowledge of how language works at the cognitive and neural level. It makes it possible to isolate language from the context of the acoustic modality, and disentangle which properties are influenced by the modality and which can be ascribed to language per se.

1.3 Comparing the temporal structure of spoken and signed languages

While there are several studies investigating the temporal structure of spoken languages both with behavioural and neurophysiological measures (Poeppel & Assaneo, 2020), sign languages have not received the same amount of research interest. Due to their intrinsic formal differences though, we can expect spoken and signed languages to have diverse temporal characteristics.

The first important difference comes from the nature of the sensory modalities employed. The acoustic modality is characterized by a much higher temporal resolution compared to the visual modality, which favours spatial over temporal information and the presentation of multiple information at the same time (Holcombe, 2009; Meier, 2002). These properties are reflected in how we experience visual and acoustic perception in our everyday life. When looking at a visual scene we can easily decompose it into several parts: shapes or sections in different locations of the scene. However, when we listen to a complex auditory input we perceive it a unified signal. For example when listening to an orchestra we might be able to pick apart the sound coming from different instruments, but overall we perceive music as a whole single input. Moreover when we look at a picture, such as a painting, the visual information is static, while acoustic perception occurs over time (i.e. music).

Modality also constraints and aids the mechanics of language production: the articulators of spoken and signed languages are very different in form and in the way they combine to produce the linguistic signal. In speech the lips, tongue, jaw and larynx are all involved in the production of sounds. The movements of multiple articulators combine to form a single perceivable acoustic signal (speech), and the visible set of articulators (mouth, lips, jaw) also function as the main oscillator for the signal (Abbs et al., 1984; Browman & Goldstein, 1992). Signed languages employ several articulators (hands, fingers, torso, head, mouth and eye-brows), which are all fully visible to the receptor of the communication; each articulator moves quasi-independently of the others and generates a different visual signal. The configuration and structure of articulators allow for parallel presentation of multiple pieces of linguistic information, whether that be lexical, syntactic or intonational (Sandler, 2018). The three phonological parameters that constitute a sign are also realized simultaneously and the way they change across time is quite different: while movement constantly changes during signing, handshape and location are more stable and less susceptible to change within a given sign. Conversely, in spoken languages information is presented, and therefore processed, in a highly sequential fashion: the speech stream is a series of phonemes that make up words.

Another difference between signed and spoken languages lies in the production rate of linguistic units. Sign duration is usually compared to syllable duration (instead of word), since most signs are considered to be monosyllabic (Coulter, 1982). Despite some discrepancy in the results due to a lack of agreement on the precise definition of sign boundaries, the findings converge to show that sign duration is about twice the duration of a monosyllabic word (Grosjean, 1977a; Klima, E. S., & Bellugi, 1979; Wilbur, 2009). These studies found a production of ~2 signs per second, compared to ~4-5 words per second. This effect has been ascribed to the size of sign language articulators, which are much bigger compared to those of spoken language. Their size, together with the greater magnitude of movements performed, leads to longer time to complete the articulatory movement. At the same time, sentence duration seems to be very similar across spoken and signed languages (Bellugi & Fischer, 1972), driven by the fact that in sign language fewer signs are needed to convey the same semantic content due to a parallel presentation of the information.

The spatial domain, completely absent in spoken languages, has a predominant role in signed languages, and, importantly, the temporal structure of sign language plays off these spatial characteristics. Overall, signed languages tend to favour a parallel and multi-dimensional presentation of information. Spoken language does include some parallel information streams, especially in the use of co-speech gestures (Özyürek, 2014) and suprasegemental prosody, but on the whole it is closer to a single sequential signal. Both the signal and the linguistic information it contains are organized on a larger time scale in signed languages compared to spoken languages, which fits well with the different temporal resolution associated with visual and auditory perceptual systems.

1.4 Language-brain entrainment

As outlined in section 1.2, language shows quasi-periodic features reflected by frequencies in the motor dynamics of speech articulators and in the changes of sound amplitude in the speech envelope. When listening to speech our brain oscillations align with these frequencies, which may support language processing and, ultimately, understanding. Entrainment consists of the temporal alignment between the oscillations in the activity of neuronal populations and the phase of an exogenous or endogenous stimulus (Obleser & Kayser, 2019). This coupling is not constrained to any specific domain, but is a widespread mechanism employed by the brain to efficiently process external (a perceptual signal such as language) or internal (motor production, inner speech, etc) information (Lakatos et al., 2019). When this phenomenon takes place in the context of language processing we talk about language-brain entrainment.

A growing body of research has developed around language-brain entrainment in the spoken language domain (see Meyer, 2018; Poeppel & Assaneo, 2020 for reviews on the topic). These studies exploit the high temporal resolution of neurophysiological measures such as electroencephalography (EEG) and magnetoencephalography (MEG) to investigate and characterize this mechanism. Language-brain entrainment has been reported in two main frequency bands: theta (4-8 Hz) and delta (< 4 Hz) (Ding & Simon, 2012; Kiebel et al., 2008; Luo & Poeppel, 2007). Theta rhythm is often associated with the syllabic modulation in the speech signal (Ding, Melloni, et al., 2017), since syllable production rate falls in this frequency range. Delta rhythm, on the other hand, encompasses slower modulations and has been linked to prosodic modulation of word and phrases (Bourguignon et al., 2013; Keitel et al., 2017; Molinaro & Lizarazu, 2018). More recently, research has started to focus on the visual component of speech, namely mouth movements. These studies investigated brain entrainment to lip movement during silent speech (in the absence of the acoustic input) showing increased synchronization in early visual cortices to frequencies matching articulatory lip movements (Bourguignon et al., 2020; Hauswald et al., 2018). These results provide evidence that entrainment takes place in the visual domain as well.

The causal role of language-brain entrainment in speech comprehension is still an open debate. Several studies found a correlation between entrainment and intelligibility (Abrams et al., 2009; Cutini et al., 2016; Doelling et al., 2014; Gross et al., 2013; Peelle et al., 2013), supporting the idea that entrainment plays an essential role in language processing. According to this idea, through entrainment the brain tracks and processes meaningful linguistic units (for review, see Giraud & Poeppel, 2012; Kösem & van Wassenhove, 2017). Conversely, other authors view entrainment as purely low-level auditory encoding, independent of the linguistic content of the signal. This view is corroborated by studies showing entrainment in several domains outside of language (Lakatos et al., 2019) and the observation that the some of the frequencies linked with linguistic units are present in the auditory cortex at rest (Giraud et al., 2007). Finally, one criticism often raised towards the language-brain entrainment studies is the use of stimuli artificially created or manipulated in the lab to present a rhythmic amplitude modulation (Alexandrou et al., 2020), which might be generating the effect.

Research on entrainment in signed language is quite scarce, probably due to some intrinsic methodological limits. The study of language-brain entrainment relies on being able to measure and characterize the temporal spectrum of the perceptual signal, and to relate it to brain activity. This task, quite easy to perform in the acoustic domain thanks to the measurement of the speech envelope, is not so trivial in the visual domain. In sign language changes over time in the perceptual signal are due to the movement of different articulators instantiating language. Researchers face the technical challenge of how to precisely measure and characterize the temporal properties of these movements.

To our knowledge only one published study tried to investigate language-brain entrainment with sign language. Brookshire and colleagues (2017) presented videos of storytelling in ASL to experienced signers and non-signers undergoing EEG. The authors developed a measure of visual signal change called Instantaneous Visual Change (IVC). IVC is based on an algorithm which calculates the differences in RGB values of each pixel across sequential frames of a video, and the sum of these squared differences gives a single metric describing the magnitude of overall change between two video frames. Calculating this value for each adjacent pair of frames yields a time series describing the visual change across the whole video. In their study, IVC was used to identify the relevant frequencies in the sign language visual signal, mirroring the function of the speech envelope in spoken language. The authors calculated coherence measures between cortical oscillations and IVC and the results showed coherence in occipital channels at 0.8-5 Hz and frontal channels at 0.5-1.25 Hz. Interestingly, no significant difference was found between signers and non-signers in occipital channels, suggesting that both groups of participants entrain equally at the level of visual perception of movement present in the videos. Conversely, coherence measures in the frontal cortex were significantly higher for signers, suggesting top-down control based on ASL knowledge.

This study represents a valuable first approach to the complex problem of investigating brain entrainment to sign language, but has an important limitation related to the measure of visual change employed. IVC measures the global visual change in a frame, discarding the specific contribution of different articulators. The perceptual and linguistic information carried by each articulator is averaged and therefore lost. The multidimensional nature of sign language calls for a better characterization of the signal, taking into account both the temporal and spatial components. In Chapter 3 I present the measure of visual signal change that we use for the experiment on language-brain entrainment with spoken and signed languages (Chapter 4). With this study I aim to clarify whether language entrainment is restricted to the acoustic domain or represents a modality-general mechanism recruited for language processing that is flexible enough to adapt to modality-specific properties of the signal.

1.5 Research questions

The main aim of this thesis is to study that role that modality plays in shaping the temporal structure of language. We tackle the effect of modality on temporal structure from two different but complementary views: the effects on the linguistic signal that is produced and on the processing of that signal by our cognitive system. The comparison of spoken and sign languages, which employ the acoustic and visual modality, respectively, represents the perfect opportunity to study this modulation while controlling for language structure. This section lays out the general research questions addressed by the studies in this thesis.

1. How does modality shape the temporal structure of the signal during language production?

The modality we use to produce the linguistic signal impacts the temporal structure of that signal itself. From the literature we find ample evidence describing differences between spoken and signed languages (see section 1.3 of this Chapter) and throughout this thesis we link these differences with modality properties. The physical properties of the acoustic speech signal have been extensively investigated by the fields of acoustic phonetics and speech and hearing science, while the visual signal of sign language is still highly understudied. In Chapter 3 we specifically focus on measuring and characterizing the spatiotemporal properties of the visual signal for both signed and spoken languages. With this study we target two specific questions:

- Can we distinguish the temporal patterns of the visual signal in sign language and spoken language? If so, which kinematic properties are better suited to describe the signal?
- Do different sign language articulators show distinct temporal patterns?

2. How is the temporal structure of the signal perceived and processed by our cognitive system?

In addition to the characteristics of the signal itself, we investigate whether language temporal structure influences the way we understand language, and how temporal distortion affects understanding. To answer this question we conduct the study presented in Chapter 2, where language intelligibility is investigated as a function of different degrees of temporal distortion of the linguistic signal. This study addresses the following questions:

- Does the temporal structure of language play a role in language comprehension?
- Does modality affect this relationship between temporal structure and language comprehension?

The main aim of this thesis is to examine the phenomenon of language-brain entrainment in the context of a signed language. With the study presented in Chapter 4 we investigate whether the brain exploits the temporal periodicity in the linguistic signal when that signal is purely visual.

- Is language-brain entrainment recruited for sign language processing?
- Does the specific temporal structure of sign language modulate the characteristics of entrainment?

This set of research questions encompass the overall theoretical framework of this thesis. In the following experimental chapters we present specific predictions and hypotheses referring to each study.

1.6 Structure of the thesis

This thesis is structured in five chapters: this General Introduction, three experimental chapters, and one final chapter presenting General Discussion and Conclusions.

The current chapter has provided the theoretical background and motivation for the studies described in the following chapters. We started with an overview of the temporal structure of language and its fundamental role in language processing (section 1.1). The majority of literature on this topic focuses primarily on spoken language, while in this thesis we systematically compare spoken and sign languages. For this reason we presented a general introduction to sign language (section 1.2) and a description of the various differences between temporal structure in sign and spoken language (section 1.3). In section 1.4 we introduced the phenomenon of language-brain entrainment, which will be the focus of the main study of this thesis (presented in Chapter 4). Finally, we defined the main research questions that this work addresses (section 1.5).

Chapter 2 presents a behavioural study investigating the effect of temporal distortion on language comprehension in different modalities. We applied a specific manipulation, called *locally time-reversed speech paradigm* (described in section 2.1), to three different types of signal (unpredictable sentences in Spanish, unpredictable sentences in LSE and a non-linguistic visual signal) to disentangle the specific contribution of signal modality and the contribution of linguistic status. The results (2.4) and discussion (2.5) sections compare, on the one hand, Spanish and LSE, and, on the other, LSE and non-linguistic visual signal to tease apart modality and language structure.

In Chapter 3 I examine the physical properties of the visual linguistic signal in both spoken and sign language. Section 3.1 explains the importance of kinematic analysis for the visual sign language signal, and different types of motion tracking systems are described in section 3.2. For this project, motion tracking data of body movements in Spanish, Russian, LSE and Russian Sign Language (RSL) were recorded with a custom-made Kinect system (section 3.3). Section 3.4 describes the different kinematic measures used to assess whether we can distinguish the visual signal in spoken and signed languages: submovements, rhythmicity, motion space, motion magnitude, and UMAP clustering applied to time-frequency data. The results of this analysis for the different languages are presented in section 3.5, and the chapter concludes with a discussion of these findings (section 3.6).

The study presented in Chapter 4 investigates the phenomenon of language-brain entrainment both in spoken and signed languages. In section 4.1 I present previous literature on entrainment in the visual modality and the effect of language knowledge on this phenomenon. A detailed description of the experiment is presented in section 4.2: we recorded the neurophysiological activity of two groups of hearing participants, expert signers and sign-naive individuals, while they were watching videos of storytelling a known spoken language (Spanish), unknown spoken language (Russian), known sign language (LSE) and unknown sign language (RSL). Next we present the analysis performed (section 4.3) and the results (section 4.4). Discussion and conclusion of this study are presented in sections 4.5 and 4.6, respectively.

The last chapter of this thesis starts with an overall summary of the results from the three studies previously presented (section 5.1). We continue by revisiting the results from each study and highlighting their contribution to the understanding of two important aspects of language processing: production of the linguistic signal and processing of that signal by our cognitive system (section 5.2). In section 5.3 we identify some limitations in our studies and suggest future research steps to clarify unsolved questions. Finally in section 5.4 we present the main conclusions of this doctoral thesis by revisiting the research questions formulated in section 1.5 of this chapter. To conclude we tie our results with potential applications in the field of sign language automatic translation.

Chapter 2: Language modality and temporal structure impact processing: sign and speech have different windows of integration

The current chapter investigates how the temporal structure of the language signal shapes the way we understand that signal. In this study we use a specific type of temporal manipulation, called *locally time-reversed speech paradigm*, which allows us to tap into the perceptual processing of different visual and acoustic signals. This experiment investigates sign language processing, and isolates, on the one hand, the specific contribution of signal modality through a comparison with spoken Spanish and, on the other hand, the contribution of linguistic status by comparing LSE with a non-linguistic temporally structured visual signal.

The study reported in this chapter was published as Rivolta, Costello & Carreiras (2021). The data, analysis scripts and example stimuli of this experiment are available in the following repository: https://osf.io/qr38u

2.1 Locally time-reversed speech paradigm

One way to investigate the role of temporal structure in language comprehension is to examine to what degree temporal distortion of the speech signal can impair intelligibility (for different techniques that employ this method, see section 1.1). The locally time-reversed speech paradigm has been employed in several experiments with different spoken languages (Greenberg & Arai, 2001; Steffen, A., & Werani, 1994; Stilp, Kiefte, Alexander, & Kluender, 2010). In this paradigm the linguistic signal is first divided into windows of fixed duration and then each window is reversed while the global order of the windows is kept. Figure 1 shows an example of this manipulation applied to speech and to Spanish Sign Language (LSE). Participants hear (or view) sentences distorted in this manner and repeat what they have understood. How intelligible the distorted signal is depends on the size of the reversal window: larger windows create higher levels of distortion and are associated with lower accuracy in the repetition task. A meta-analysis of several studies employing this paradigm (Ueda, Nakajima, Ellermeier, & Kattner, 2017) showed that the intelligibility pattern is consistent across different spoken languages, even when they are characterized by different timing patterns (e.g. syllable-, mora- or stressed-based). Intelligibility starts to decrease when the reversal window is approximately 40 ms, drops under 50% when the window is between 60-70 ms and for windows of 100 ms or longer speech comprehension is almost completely lost.

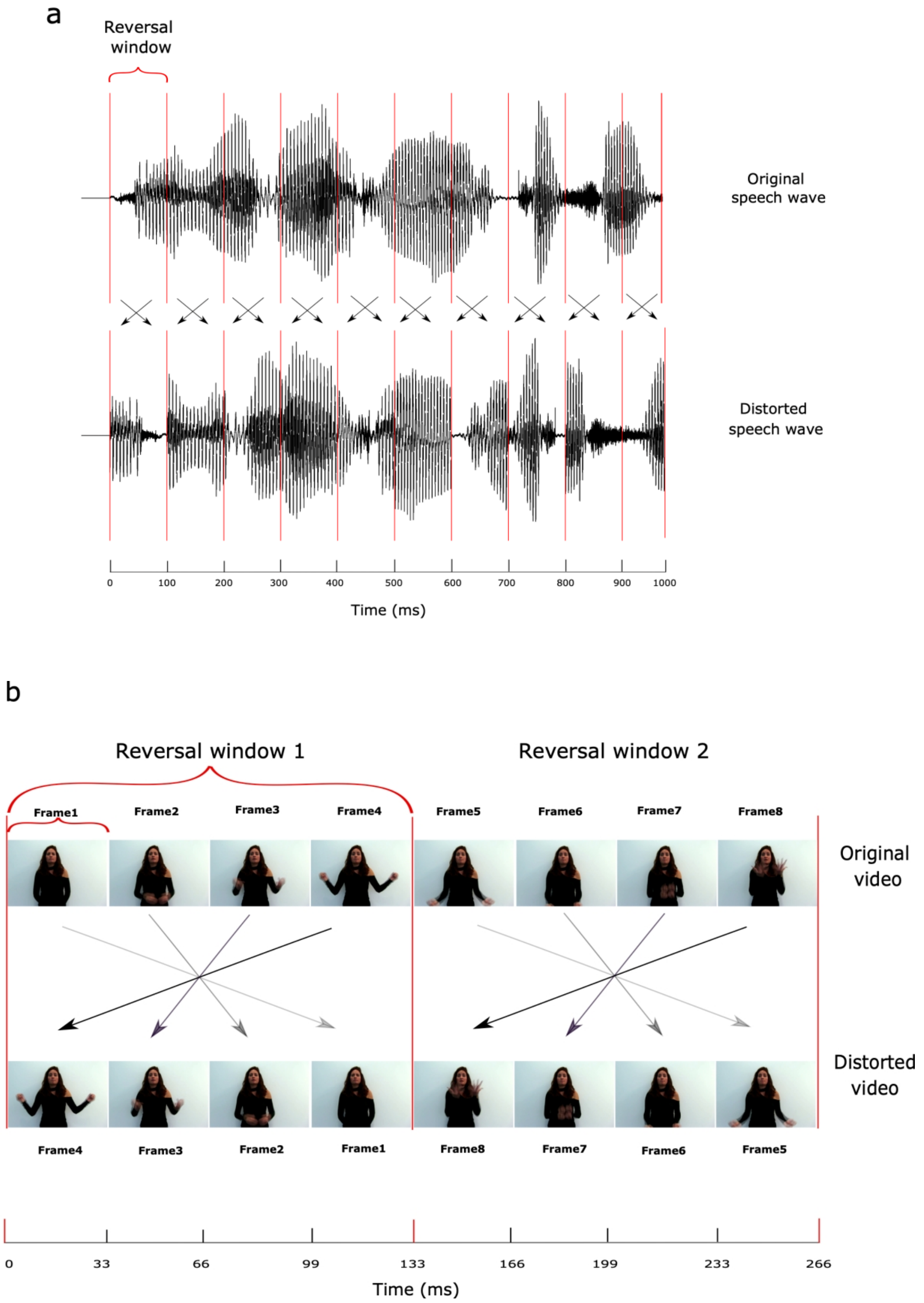

Figure 1: A schematic to illustrate locally time-reversed speech paradigm applied to (a) Spanish and (b) LSE. In Spanish (a) the acoustic signal from the sentence 'El premio normal bloquea el tiempo preciso' ('The normal prize blocks the precise time') is divided into reversal windows of 100 ms and the signal is reversed inside each window, while the order of the windows themselves is maintained. In LSE (b) the video is made up of frames that last 33 ms each; the video is divided in reversal windows of 133 ms (4 frames),

and the order of the frames is reversed within each window. The schematic shows the frames of the first sign, MILAGRO *['miracle'], of a sentence.*

To our knowledge only one study has adapted this paradigm to a sign language. Hwang (Hwang, 2011) investigated the locally time-reversed paradigm with deaf signers (native and late-learners) of American Sign Language (ASL). The results confirmed the general tendency of language intelligibility to decrease with longer reversal windows, but also revealed important differences from the pattern associated with spoken languages. Firstly, the temporal scale of the signed signal is slower than that of speech, and this was factored into the experimental design by using larger reversal windows for sign language. Intelligibility of the ASL stimuli was characterized by a slow, constant decrease as the size of the reversal window increased. The decrease stops at reversal windows of approximately 500 ms, at which point intelligibility starts to level off at 50% accuracy. A comparison between native and late learners of ASL showed an effect of age of acquisition: the intelligibility level of late-learners was lower than that of native signers across all reversal windows. Nevertheless, the intelligibility decrease showed the same pattern, namely, an initial decline which plateaued at around 50% accuracy level, for both groups.

2.2 The experiment

In this study we use the locally time-reversed speech paradigm to investigate how our cognitive system temporally processes the incoming linguistic signal and how temporal distortion affects sign language comprehension. We are interested in disentangling the impact that signal modality and linguistic features may play in language decoding. The design had three separate tasks, each with a different type of material: spoken Spanish sentences, LSE sentences and non-linguistic visual stimuli. This design lends itself to making two types of comparisons. Firstly, a comparison between spoken and signed language, in this case Spanish and LSE, can reveal how acoustic and visual perceptual properties modulate the temporal processing of language. Secondly, we want to compare sign language with non-linguistic visual material to investigate to what extent the results found in LSE are due to specific linguistic properties or general principles of the visual modality.

We selected non-linguistic stimuli (see sections 2.3.2 for detailed explanations of the stimuli) that were as comparable as possible to language in terms of regularity in their temporal structure; importantly, the non-linguistic stimuli had no phonological or syntactic structure. The same manipulation of local time reversal was applied to all three stimulus types, but the paradigm was slightly adapted according to the stimulus type. Reversal windows were different depending on the modality, and were chosen based on the results of previous studies employing this paradigm in spoken (Greenberg & Arai, 2001; Kiss, Cristescu, Fink, & Wittmann, 2008; Ueda et al., 2017) and signed (Hwang, 2011) languages. The range of reversal window sizes was balanced to detect both the intelligibility threshold and possible plateau effects at the extremes of the intelligibility curve. In line with the differences in temporal resolution between the auditory and visual modalities described above, reversal windows in the visual modality were larger and increased in larger steps than those used for spoken language.

If language information is temporally decoded in order to allow comprehension, and all types of linguistic signal rely on the same underlying temporal structure, we would expect to find similar patterns for both Spanish and LSE: a marked breakdown in intelligibility at a specific threshold. However, given the differences in the pseudo-periodicity of each type of signal (speech is a relatively fast changing signal compared to sign language) the threshold for intelligibility, which is typically around 40 ms for speech, should be at windows of a much longer duration for sign language. Conversely, different modalities may exploit temporal structure in qualitatively different ways: in this case, we expect Spanish and LSE to be differently affected by the temporal manipulation. The previous results found by Hwang (Hwang, 2011) suggest that this is the case. If signed languages are indeed characterized by a common temporal structure, and furthermore this temporal organization differs from that of spoken languages, we should find similar results when testing another sign language: LSE (Spanish Sign Language). Finally, if temporal structure is easier to segment when the input is linguistic, we should find better performance in LSE. In contrast, if the visual signal is parsed in the same way and information extracted similarly when processing visual material, we should not find differences when comparing the LSE with a non-linguistic task.

The study consisted of a single session with three experimental tasks: spoken Spanish, LSE and visual non-linguistic. The general characteristics of the study, with separate sections for each task to describe the materials and procedure, are presented in the following section (section 2.2).

Afterwards we present the results of each task analysis and comparisons across different tasks (section 2.3) and we discuss the findings (section 2.4).

2.3 Methods

The experiment was conducted in different locations across Spain with the same equipment. The experiment ran on a DELL portable computer with Windows 7 OS, using Psychopy (version 1.85.3) in Python (version 2.7.11). All participants heard acoustic stimuli through headphones at the same comfortable volume. Responses in Spanish and LSE were recorded with a video camera. The order of the tasks was randomized across participants. All participants signed an informed consent form before the beginning of the experiment and were compensated for their participation.

The research was conducted with prior approval of BCBL Ethics Review Board and complied with the guidelines of the Helsinki Declaration.

2.3.1 Task 1: Spanish

2.3.1.1 Participants

Twenty-three participants took part in the experiment, 18 females and 5 males, with a mean age of 40 (29 - 51 years). All participants were bimodal bilinguals who were native speakers of Spanish and native or highly proficient users of LSE. Proficiency in LSE was reflected in self-reported ratings on a likert scale from 1 (no knowledge) to 5 (very good knowledge), with a mean rating of 4.81 (SD= 0.40) across all participants. Four participants were native signers, as they learned LSE before 1 years of age from a family member; those participants who were not native LSE signers were professional sign language interpreters.

2.3.1.2 Material

Stimuli consisted of 60 semantically unpredictable sentences in European Spanish, which were syntactically and grammatically correct sentences with no sensible meaning. The use of semantically unpredictable sentences ensures that the results are due to correct perception of the sound and not to inference based on pragmatics or linguistic context (Greenberg & Arai, 2001; Hwang, 2011). We generated the sentences with a set of 120 adjectives, 120 nouns and 60 verbs.

We selected the words from the European Spanish subtitle corpus of the EsPal database (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) controlling for frequency (log_count between 3-7), number of phonemes per word (between 3-10) and number of syllables per word (between 2-4). Because the same words were used to generate the Spanish and LSE sentences (see section 2.2.2), words that were homonyms in LSE were excluded, as were those which corresponded to compound signs or classifier-based signs with a very generic meaning.

The words were then randomly combined together to create 60 sentences with the following structure:

Determiner$_1$ + Noun$_1$ + Adjective$_1$ + Verb + Determiner$_2$ + Noun$_2$ + Adjective$_2$.

See Table 1 for example sentences. No noun, adjective or verb was repeated across sentences. A native speaker reviewed all the sentences to make sure that they were grammatically correct.

Table 1: Examples of semantically unpredictable sentences used in the experiment.

Spanish sentences	English translation
La luna urgente persigue al diablo bueno.	The urgent moon chases the good devil.
El interés ácido despide al cuchillo cojo.	The acid interest dismisses the lame knife.
El pueblo invisible divide el monstruo blanco.	The invisible village divides the white monster.
El campamento caro aguanta el puente azul.	The expensive camp holds the blue bridge.
El teatro contrario adora la distancia furiosa.	The opposite theatre adores the furious distance.
El empleo serio contiene la habilidad oscura.	The serious job contains the dark skill.
La mitad directa solicita la pasta negra.	The direct half requests the black pasta.
La planta favorita revisa la sangre fría.	The favourite plant checks the cold blood.
El vestido curioso demuestra el cuento vago.	The curious dress shows the vague tale.

A female native speaker recorded 65 sentences (60 experimental sentences and 5 practice sentences) with natural prosody and pace, using a Sennheiser ME65 microphone in a sound-proof recording booth. The item duration varied between 2.29 and 2.98 s (mean = 2.65 s, SD = 0.16). The recordings were normalized for sound level (70dB) and distorted with the reversal window manipulation using Praat (Boersma, P., & Weenink, 2020). Five reversal windows were applied: 40 ms, 55 ms, 70 ms, 85 ms and 100 ms. At the beginning of each distorted sentence 50 ms of silence was added to avoid clipping due to the loading time of the audio file in the experiment presentation software.

2.3.1.3 Procedure

Participants heard 60 items, each in one of six conditions: undistorted or distorted with one of the five reversal windows. Items were pseudo-randomly assigned to different conditions across participants such that each participant heard ten items in each condition, and across all participants each item appeared the same number of times in each condition.

Participants read instructions in Spanish and then performed five practice items before beginning with the experimental trials. After hearing an item, participants could choose to respond or to hear the same item again three more times (up to a total of four presentations for the same item) before giving their answer (following Greenberg & Arai, 2001; Hwang, 2011). Participants gave their response by repeating out loud the sentence they believed that they had heard. Responses were scored for word identity and word order. For word identity participants received one point for each lexical word that was correctly produced, up to a maximum of five points (given that there were five open class words in each sentence). For word order participants received up to four points for the correct relative order between each pair of lexical words in the sentence. The maximum score for each sentence was of nine points.

2.3.2 Task 2: Spanish Sign Language

2.3.2.1 Participants

Participants were the same as those for Task 1 (see section 2.1.1).

2.3.2.2 Material

The material consisted of 60 semantically unpredictable sentences in Spanish Sign Language (LSE). We generated the sentences with the same sets of words used for the Spanish sentences (see section 2.1.2) by recombining the words together with the same random procedure. Each word was translated into the corresponding LSE sign taken from the Standard LSE Dictionary (Fundación CNSE, 2008) to avoid regional variants. All the signs had unique forms and none was a compound (to avoid phonological complexity) or a classifier-based sign with a generic meaning (to avoid semantic ambiguity). It was not possible to control for sign frequency and other lexical properties because this information is not available for LSE.

Following the grammar and syntax of LSE (Herrero Blanco, 2009) all sentences were recorded with the following SOV structure:

Noun$_1$ *+ Adjective*$_1$ *+ Noun*$_2$ *+ Adjective*$_2$ *+ Verb.*

No determiner sign was present in the sentences since LSE marks this feature by other means.

A deaf female native signer modeled 65 sentences in LSE (60 experimental sentences and 5 practice sentences) with natural prosody and pace, recorded at 25 fps with a video camera (Sony HDR-CX240E). The model signed in front of a uniform white background and each sentence began and ended with hands in a resting position in front of the body.

The videos of LSE sentences were pre-processed with FFmpeg (Tomar, 2006) (version 2.7): each video was cropped (810 x 540 pixels), the frame rate was set to 30 fps to match the 60 Hz refresh rate of the presentation screen, and the luminance was normalized across videos. The item duration varied between 3.36 and 5.76 s (mean = 4.67 s, SD = 0.48). The reversal-window manipulation was applied using a custom Python script and a fade in/out of 3 frames (100 ms) was applied to each of the manipulated sentences using FFmpeg. The reversal window sizes for LSE sentences were: 4 frames (133 ms), 6 frames (199 ms), 8 frames (266 ms), 10 frames (333 ms) and 12 frames (399 ms).

2.3.2.3 Procedure

The procedure was the same as for Task 1, except that the participants saw the LSE stimuli on screen and gave their responses by signing.

2.3.3 Task 3: Visual non-linguistic stimuli

2.3.3.1 Participants

Participants were the same as those for Task 1 (see section 2.1.1).

2.3.3.2 Material

The material consisted of 36 videos (30 experimental videos and 6 practice videos) representing four symbols traced one after the other in the center of screen by a moving dot that left no line. The symbols used were six digits (1, 2, 3, 6, 8, 9) and six letters (C, L, O, S, V, Z). We chose to use letters and digits as symbols because they represent very well-known shapes that can be identified with a button press on a keyboard. Even if letters and digits have linguistic labels that may have been used to encode the elements in memory, the signal that presented the sequence is not language-like: the elements had no structure either internally (phonology) or in relation to each other (syntax). The pattern of the dot tracing the symbols was modeled after natural handwriting using a custom Matlab (version 2018b) script; all the symbols were written in one continuous stroke and were easily recognizable.

Each video consisted of a fixed sequence of two letters followed by two digits. The pool of 12 symbols was randomly combined together to fit this sequence, so that a symbol never appeared twice in a given trial and each symbol appeared the same number of times across all stimuli.

All videos were then manipulated using a custom Python script with the same procedure and the same reversal windows used for the sign language sentences: 4 frames (133 ms), 6 frames (199 ms), 8 frames (266 ms), 10 frames (333 ms) and 12 frames (399 ms)

2.3.3.3 Procedure

The procedure was similar to the one used in Task 2. Each participant saw each of the 30 videos in one of the six conditions (undistorted or with one of the five reversal windows). The condition in which a given item appeared was counterbalanced across participants. In contrast with the linguistic tasks, here participants could see each video only once before responding by typing the symbols they recognized on the keyboard. The difference was implemented to avoid ceiling effects; the results of a pilot session revealed high accuracy rates in Task 3 due in large part to the small set of stimuli used in the task. The instructions informed participants that the sequences consisted of two letters followed by two digits. Responses were scored by symbol identity (up to four points, one for each symbol) and symbol order (up to two points for the relative position between the two letters and the two digits), giving a maximum score of six points for each item.

2.4. Results

All statistical analyses for the experimental tasks were run using R (R Core Team, 2017) (version 3.6.2); mixed linear models were run using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) and analyzed with the lmerTest package (Kuznetsova, A., Brockhoff, P. B., & Christensen, 2013).

2.4.1 Analysis of sentence repetition

In both linguistic tasks participants could decide to hear or view each sentence up to four times before reproducing it. The distribution of the number of sentence presentations across reversal windows is different for Spanish and LSE, as shown in Figure 2. While in Spanish participants requested more presentation of sentences with longer reversal windows, in LSE participants tended to view each sentence multiple times even when the reversal window was very short or in the (undistorted) baseline condition (we present a possible explanation for this strategy in section 4.1).

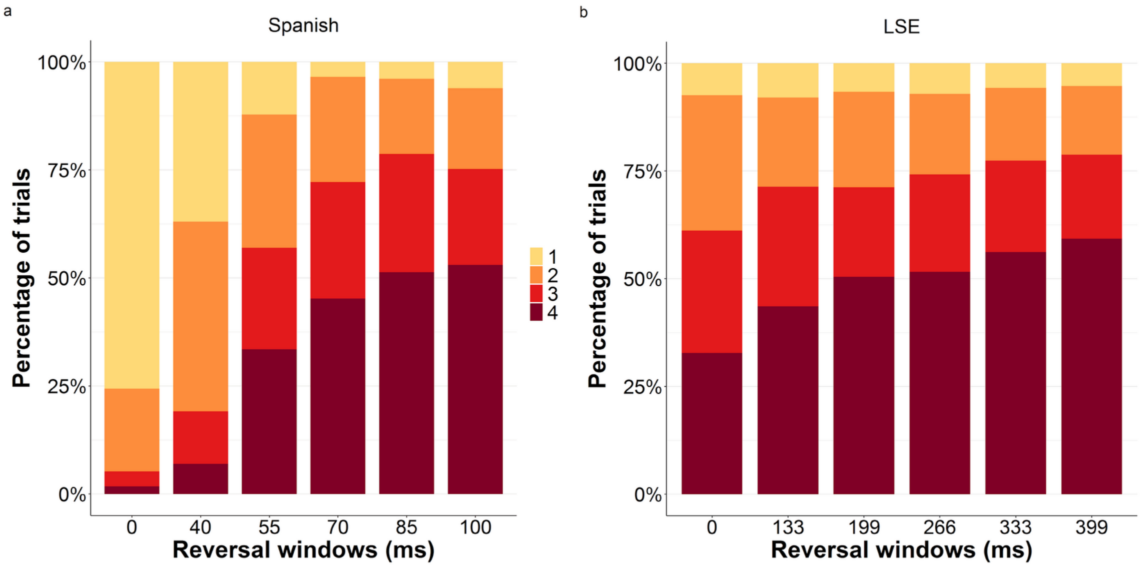

Figure 2: Distribution of the number of stimulus presentations across reversal windows in (a) Spanish and (b) LSE.

In order to investigate whether the number of sentence presentations modulates the intelligibility of the sentence we ran two separate linear mixed models for each language. Accuracy represented the dependent variable, while Reversal window and Number of presentations were input as continuous predictors; the random effects were intercepts for participant and item and a by-participant random slope for the effect of reversal window. The Spanish model (R^2=.77) showed a statistically significant effect for reversal window (β= -14.22, *SE*= 1.77, *p* < .001), number of presentations (β= -6.22, *SE*= 0.77, *p* < .001) and the interaction between the two (β= -8.17, *SE*= 0.6337, *p* < .001). For Spanish sentences, intelligibility decreases with longer reversal windows and with a higher number of presentations. Lower accuracy scores were associated with longer reversal windows and, counterintuitively, with a higher number of sentence presentations: highly distorted sentences were more difficult to understand and therefore participants tended to hear them more times. The interaction effect is driven by the fact that a higher number of sentence presentations was associated with better performance only in the longer reversal windows: in shorter reversal windows, the pattern was reversed, as can be seen in Figure 3.

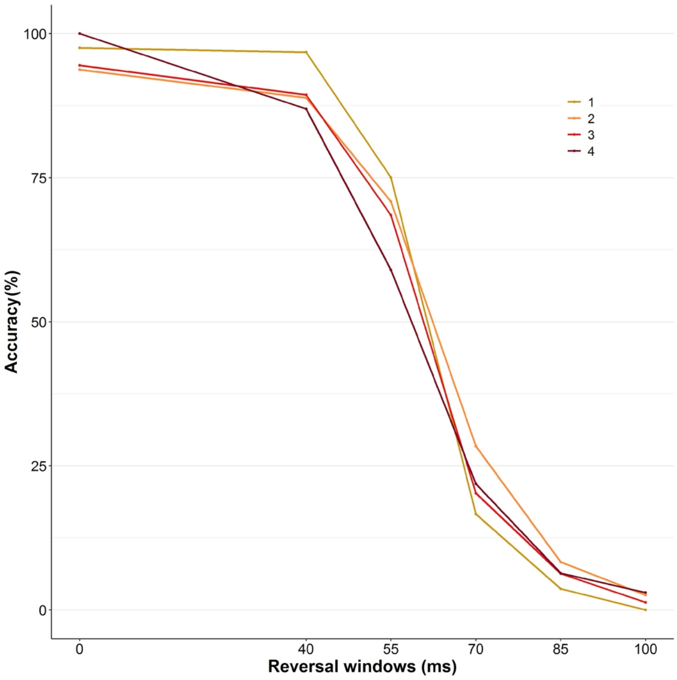

Figure 3: Accuracy as a function of reversal window size and number of sentence presentations in Spanish.

The LSE model (R^2=.49) showed a statistical significant effect for reversal windows (β= -11.52, *SE*= 0.70, *p* < .001), but the main effect of number of presentations (β= -0.65 , *SE*= 0.78 , *p* = 0.39) and the interaction (β= 0.55 , *SE*= 0.65 , *p* = 0.39) were not significant. In LSE participants viewed each sentence multiple times independently of the level of temporal distortion.

Number of sentence presentations was not included in the following analyses, where intelligibility is compared across different tasks.

2.4.2 Comparison between Spanish and LSE

To compare the results in LSE and Spanish we treated reversal windows in both languages as a categorical variable, making it possible to match the different absolute values of the windows used in each language (for a justification of reversal windows sizes see section 2.2). We ran a linear mixed model including Reversal window and Task as categorical predictors; intercepts for participants and items were input as random effects. Main effects and interactions were assessed by calculating Type III F-statistics and significance p-values using Satterthwaite approximations to denominator degrees of freedom (Casaponsa et al., 2019). The model (R^2=.70) showed a statistically significant effect for both Reversal window ($F_{(5, 2658)}$ = 792.02, p < .001) and Task ($F_{(1, 2657.1)}$ = 441.48, p < .001). Moreover Reversal window and Task showed a significant interaction in modulating intelligibility ($F_{(5, 2658)}$ = 265.61, p < .001). We performed a post hoc analysis on the model fitted values, comparing all consecutive reversal windows (rev1-rev2, rev2-rev3, rev3-rev4, rev4-rev5 and rev5-rev6) in each task and paired reversal windows across tasks. Consecutive post-hoc comparisons, corrected using the Bonferroni method, are shown in Table 2. Post-hoc comparisons across tasks are shown in Figure 4 (for detailed values of the post-hoc analysis see Table A1 in Appendix 1).

In both tasks larger reversal windows were associated with lower intelligibility, but the pattern is different across the two languages. In Spanish temporal distortion with the smallest reversal window (40 ms) does not create any detrimental effect but intelligibility decreases sharply between window sizes of 40 and 85 ms, reaching almost complete loss of word recognition. The results for LSE show a gradual decrease in intelligibility as the size of the reversal window increases: an initial dip between 0 and 133 ms followed by a less pronounced slope. For LSE, intelligibility scores spanned a limited range: in the baseline condition (0ms) average intelligibility only reached 86% and even in the longest reversal window condition (399 ms) it never drops below 50%.

Table 2: Summary of results of post-hoc t-tests (corrected using the Bonferroni method) comparing intelligibility between consecutive reversal windows for the Spanish and LSE stimuli. Post-hoc and descriptive statistics (mean and SD) are based on fitted data from the linear mixed model comparing the tasks in Spanish and LSE.

Reversal window	*Spanish*					*LSE*				
	Reversal window duration (ms)	Accuracy (%)		Comparison (with next window)		Reversal window duration (ms)	Accuracy (%)		Comparison (with next window)	
		Mean	SD	*t*	*p*		Mean	SD	*t*	*p*
1	0	96.9	2.26	2.67	0.12	0	86.6	2.27	9.45	<.001
2	40	91.7	2.26	12.85	<.001	133	68.3	2.27	0.62	1.00
3	55	66.8	2.26	22.77	<.001	199	67.1	2.27	3.43	0.01
4	70	22.7	2.26	8.39	<.001	266	60.3	2.27	3.28	0.02
5	85	6.46	2.26	2.16	0.48	333	53.9	2.27	1.66	1.00
6	100	2.27	2.26			399	50.6	2.27		

The comparison across paired reversal windows can inform us how increasing degrees of temporal distortion affect Spanish and LSE (Figure 4). In the two smallest reversal windows intelligibility was higher in Spanish compared to LSE; in the three largest windows this pattern inverts as intelligibility in LSE was significantly higher than that of Spanish. Although intelligibility for undistorted sentences in LSE never reached 100%, these results reveal that sign language is more resilient than spoken language to this type of temporal manipulation: even under highly distorted conditions participants were able to recognize and repeat about half of the signs presented in the sentence. We return to these issues in the general discussion in section 2.5.1.

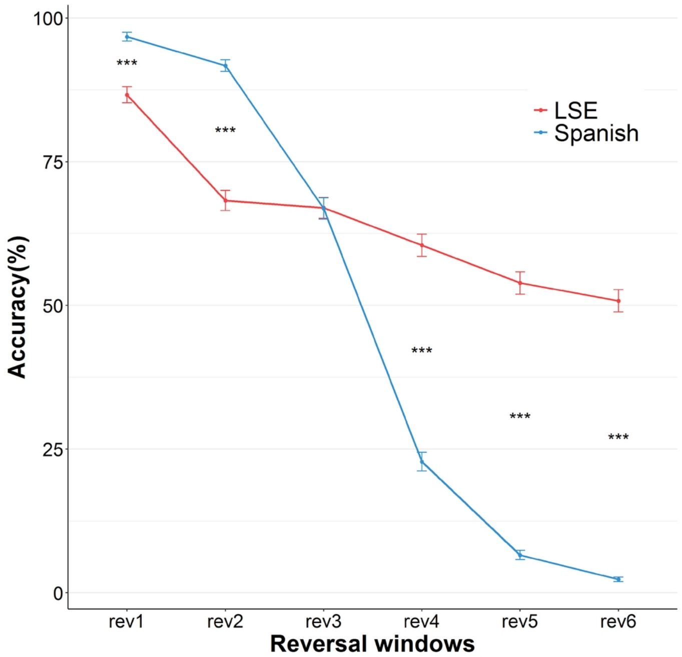

*Figure 4: Accuracy as a function of reversal window for the Spanish task (blue) and LSE task (red). Points joined by lines show intelligibility averaged across participants (n = 23) in each task, while error bars show the standard error of the mean for each point. Asterisks show statistically significant post-hoc comparisons across tasks in matched reversal windows (*** p < .001).*

2.4.3 Comparison between LSE and the visual non-linguistic task

To compare results in LSE and the visual non-linguistic task we carried out the same analysis we used to compare Spanish and LSE above. The model (R^2=.53) showed a statistically significant effect for Reversal window ($F_{(5,\ 1965.9)}$ = 237.07, *p* < .001), Task ($F_{(1,\ 2016,2)}$ = 190,57, *p* < .001) and their interaction ($F_{(5,\ 1965.9)}$ = 33,48, *p* < .001) in modulating intelligibility.

We performed post hoc analysis on the model fitted value, comparing all consecutive reversal windows (rev1-rev2, rev2-rev3, rev3-rev4, rev4-rev5 and rev5-rev6) in each task and paired reversal windows across tasks. Consecutive post-hoc comparisons, corrected with Bonferroni, are shown in Table 3; comparisons across tasks are shown in Figure 5 (for detailed values of the post-hoc analysis see Table A2 in Appendix 1). A trend of lower accuracy with larger reversal windows was common to both tasks, with some differences.

Table 3: Summary of results of post-hoc t-tests (corrected using the Bonferroni method) comparing intelligibility between consecutive reversal windows for the visual non-linguistic stimuli and LSE stimuli. Post-hoc and descriptive statistics (mean and SD) are based on fitted data from the linear mixed model

comparing the visual non-linguistic and LSE tasks.

Reversal window	***Non-linguistic stimuli***					***LSE***				
	Reversal window duration (ms)	**Accuracy (%)**		**Comparison (with next window)**		**Reversa l window duratio n (ms)**	**Accuracy (%)**		**Comparison (with next window)**	
		Mean	SD	*t*	*p*		Mean	SD	*t*	*p*
1	0	87.7	3.39	6.03	<.001	0	86.6	3.01	8.72	<.001
2	133	69.8	3.39	5.85	<.001	133	68.3	3.01	0.52	1.00
3	199	52.4	3.39	5.71	<.001	199	67.2	3.01	3.24	0.02
4	266	35.5	3.39	3.82	<.01	266	60.3	3.01	2.95	0.05
5	333	24.1	3.39	1.64	1.00	333	54.1	3.01	1.66	1.00
6	399	19.2	3.39			399	50.5	3.01		

The analysis of the visual non-linguistic task showed that accuracy in recognizing the symbols consistently lowered as the reversal window increased from 0 to 266 ms, but this drop in accuracy stopped at 333 ms. Participants found the task challenging, as reflected by the accuracy ranging between 87.7% and 19.2%. The comparison between performance in LSE and in the visual non-linguistic task (Figure 5) reveals how accuracy is similar up until 133 ms reversal windows, but for larger windows the two curves start to diverge: the accuracy in LSE is higher than that of the visual non-linguistic task.

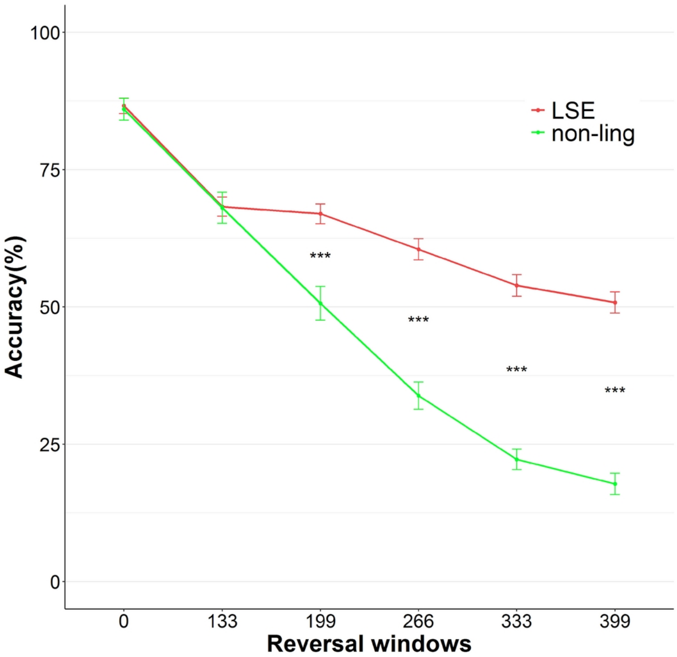

*Figure 5: Accuracy as a function of reversal window size the visual non-linguistic task (green) and sign language task (red). Points joined by lines show intelligibility averaged across participants (n = 23) in each task, while error bars show the standard error of the mean for each point. Asterisks show statistically significant post-hoc comparisons across tasks in matched reversal windows (*** p < .001).*

2.5 General discussion

The results from each task, taken individually, give insight into how temporal distortion affects the perception of different types of signal. In Spanish it is possible to identify a clear threshold (40 ms reversal window) up to which incoming acoustic information is only minimally affected by temporal distortion. Once this threshold is reached though, the ability to perceive well-formed words from the acoustic signal is rapidly and almost completely lost. These results closely reproduce the findings in the literature on locally time-reversed speech: this pattern has been reported for various spoken languages (Ueda et al., 2017). The experiment extends the results of locally time-reversed speech paradigm to spoken Spanish, and adds to the growing body of evidence that the cognitive system perceives and parses different spoken languages with a common mechanism.

LSE, on the other hand, shows a more gradual pattern: longer reversal windows elicit an increasing loss of intelligibility, but the reduction is gradual and constant across reversal windows. Additionally, the cognitive system is still able to extract enough information to identify some of the signs in the distorted visual signal. The overall pattern in LSE matches the findings from the earlier study with deaf signers of ASL: in that study, participants' accuracy gradually decreased and spanned between 90% and 50% (Hwang, 2011). That study and the current experiment represent the only two studies applying locally time-reversed paradigm to signed languages; more research is therefore needed to draw strong conclusions but the similarities between the LSE and ASL results suggest that there may be a common temporal mechanism for visual signed language.

In the following sections we compare the results across language modality (spoken vs sign) and stimulus type (linguistic vs non-linguistic) to gain a fuller picture of the role played by modality and linguistic structure in temporal processing.

2.5.1 Comparison between spoken and sign language

The main goal of this study was to investigate how language modality impacts the way our cognitive system analyses and parses the incoming linguistic signal. Spanish and LSE differ qualitatively in how intelligibility is modulated by increasingly longer reversal windows. The difference between Spanish and LSE resides in the absence of a clear drop in intelligibility corresponding to a specific reversal window size in LSE. This reversal window, sometimes defined as temporal integration window (Poeppel, 2003), represents the temporal resolution unit for spoken language processing.

In general LSE appears to be more resistant to temporal distortion than Spanish. Two factors may explain the resilience of the sign language signal. Firstly, signed languages make use of the visual domain, which relies less on temporal structure. Intuitively, this makes sense: we have little difficulty in reversing a movement, but reversing a sound is a much more onerous task. Support also comes from neuroimaging studies that have used backwards signing as a baseline condition and report that signers could identify some lexical items (Inubushi & Sakai, 2013). In a recent study Bosworth and colleagues (2020) presented ASL narratives backwards, and found that signers were still able to understand and recall a good part of the narrative. Secondly, the spatial character of sign languages means that information is maintained over time: phonological parameters, such as handshape, are stable in the signal for enough time to make it possible to recover these features even when the temporal properties are degraded. Evidence from the production of backwards signing supports this idea that groupings of phonological features remain accessible within the temporal syllabic structure of the sign (Wilbur & Petersen, 1997). Furthermore, the use of space allows different features to appear at the same time: identifying a combination of features, such as a handshape and a location, may be sufficient to identify a lexical item even when all the other parameters are unidentifiable or highly distorted. This relates to the distribution of information in the sign language lexicon and possible redundancies in the signal. This redundancy is evident also in previous work on the lexical recognition of signed and spoken languages: Emmorey and Corina (1990) found that lexical recognition is faster and easier for signs compared to spoken words and argued that this is due to spatial and temporal properties associated with sign structure.

Despite the resilience associated with signed languages, LSE performance in conditions with little or no distortion was lower than that of Spanish, and never reached ceiling accuracy. An explanation for this result might lie in differences in task difficulty. Participants were all native Spanish speakers, whereas for most of them LSE was their second language (albeit with a high level of proficiency). Language proficiency of participants might therefore account for the lower performance in undistorted sentences and the higher level of variability in participants' performance during the task in sign language. Age of acquisition does modulate performance on this type of task: when the locally time-reversed paradigm is performed by non-native speakers intelligibility starts to decline at shorter reversal windows compared to native speakers, but the general trend and shape of the intelligibility curve is similar across native and non-native participants (for spoken German, (Kiss et al., 2008); for ASL, (Hwang, 2011)). Being a non-native speaker (or signer) modulates resilience to locally time-reversed speech, but the overall pattern remains unchanged. Nevertheless, the same effect is evident in the previous ASL study (Hwang, 2011), in which deaf native signers had a baseline accuracy of approximately 90%, comparable to our result. Another independent study found similar results while testing the intelligibility of time-compressed ASL sentences in deaf native signers. In this population the accuracy for sentences played at the normal rate was at 88% (Fischer et al., 1999). The errors for undistorted LSE (and ASL) can be explained by the higher short term memory effort associated with signed languages. Short term memory span in signed language is about 5± 2, compared to the classical 7± 2 associated with auditory presented stimuli (Boutla, Supalla, Newport, & Bavelier, 2004). Different explanations have been proposed to account for this difference, such as a phonological similarity effect across signs or sign articulatory length. Recent work with second language learners of signs suggests a possible role of perceptual-motor memory processes in the sign lexicon (Martinez & Singleton, 2018). The lower baseline accuracy for sign languages appears to be related not to age of acquisition but to linguistic properties that affect processing. Future work comparing native and non-native signers could tease apart the relative contribution of proficiency, age of acquisition and modality.

In both linguistic tasks participants could hear or see the same sentence up to four times before giving their answer. In the spoken language task the interaction of number of presentations and reversal window size modulated intelligibility. More repetitions improved performance in the longer reversal windows (70 – 100 ms), while in shorter reversal windows fewer repetitions were associated with better intelligibility. In the sign language task, on the other hand, we observed a different distribution in the number of repetitions: most participants tended to watch each sentence three or four times, independently of the size of the reversal window. Participants reported struggling to retain in memory the signs from the sentence much more than they did with Spanish words; as a result they often decided to see the sentence more than once, even in the baseline condition. This strategy seems to be driven by the greater short term memory effort associated with maintaining and retrieving signed language items (Boutla et al., 2004; Hall & Bavelier, 2011). Since nearly all of our participants were not native signers, non-native processing might be driving this effect. However, data from a sentence repetition task in ASL show that (deaf and hearing) native signers also struggle with this sort of task, and, more importantly, suggest that fluency (rather than hearing status) is what modulates the type of working memory strategy employed to carry out the task (Supalla, Hauser & Bavelier, 2014). The number of repetitions of sign language sentences in our study does not modulate accuracy because participants chose to see most items as many times as possible.

Overall our results indicate that sign language does not share the same temporal resolution as spoken language, and suggest that the temporal processing of language relies on a mechanism which is at least partly modality dependent. We acknowledge some limitations in the comparison between Spanish and LSE. The manipulation applied to spoken and signed language may affect the acoustic and the visual modality differently. The locally time-reversed speech paradigm is designed to distort the temporal order of the signal. As the spatial domain plays a fundamental role in signed languages, a distortion that specifically targets its spatial organization — as opposed to the temporal one — might impair language intelligibility to greater extent. Moreover, age of acquisition and proficiency level of LSE could partly modulate their performance in the task, although our results (with hearing bimodal bilinguals) concur with those of deaf native signers (Hwang, 2011). Nevertheless, future studies employing spatial distortions and investigating a different population (native signers and less proficient signers) could contribute to a better understanding of sign language structure and processing.

2.5.2 Comparison between linguistic and non-linguistic material

Comparing results from the LSE task and the visual non-linguistic task casts light on the specific role of language properties in the temporal parsing of a visual signal. In LSE participants had to reproduce signs within a sentence, while in the non-linguistic task they saw videos of a dot tracing a sequence of four easily recognizable and nameable symbols (two letters and two digits). The sequential presentation of the symbols, and the fact that participants had to identify the symbols in the correct order, parallels the recognition of signs in the LSE task. Overall, we believe that our non-linguistic stimuli can provide a valid term of comparison, although the tasks are not identical. An important difference is the structure of the stimuli: while both types of materials rely on the spatial dimension, the multiple articulators which characterize LSE are not present in the non-linguistic task, which means that information is not presented simultaneously. Moreover, the sets of stimuli used in the two tasks are different: 300 signs in LSE compared to only 12 symbols for the non-linguistic task. The probability of presentation for each symbol is therefore much higher than that for the signs. Another difference lies in the paradigm used in the two tasks: while in LSE participants could view each sentence up to four times before giving their answer, in the visual non-linguistic task they could see each video only once. The analysis showed that the number of sentence repetitions in the LSE task had no significant effect on language intelligibility but this null effect may be due to the fact that participants chose to view most sentences four times, regardless of the degree of distortion. More critically, the two tasks showed the same levels of accuracy for the undistorted stimuli, as can be seen in Figure 5, suggesting that in the baseline condition they were comparable. The accuracy curves start to diverge only after the second reversal window and this difference between the two tasks cannot be ascribed to a simple difference in how many times the stimuli were viewed by participants.

The results show that in both tasks the accuracy in perceiving distorted stimuli gradually decreases as reversal window size increases. In contrast to the response for spoken language, neither curve has a clear point where intelligibility and recognition of the stimuli is lost, suggesting that the visual domain is not as susceptible to temporal disturbance. Although the pattern is very similar across the visual tasks, in the non-linguistic task accuracy drops more than in the LSE task: intelligibility in LSE never drops under 50% while in the non-linguistic task accuracy falls to 20%. This difference is surprising, since intuitively symbol recognition should be easier than sign recognition given the restricted set of symbols used in the experiment. Both tasks benefit from the superiority of the visual modality in managing temporal distortion, but LSE shows some additional advantage. As mentioned in the comparison of the spoken and sign language results (section 2.5.1), the spatial and temporal structure of signs makes possible an over-representation of the linguistic information in the signal. Evidence comes from studies where participants were still able to recognize signs with high accuracy even when the information in the signal was reduced, for example when the videos were presented with fewer frames per second (Johnson & Caird, 1996) or speeded up by a factor of three (Fischer, Delhorne, & Reed, 1999). The relative resilience to temporal distortion of LSE with respect to the non-linguistic stimuli suggests that features of the linguistic signal in this modality aid recognition of temporally distorted signs. These features may include the simultaneous articulation afforded by the visuo-spatial channel as well as the combinatorial properties of sub-lexical units of sign language.

The results point toward a common mechanism for the temporal resolution of visually-presented stimuli that is characterized by a constant reduction in accuracy as the signal becomes more distorted temporally. The organization of visual information in signed languages appears to attenuate how much information is lost by any disturbance of the temporal structure.

2.6 Conclusion

The fundamental role that temporal structure plays in speech comprehension calls for a clear characterization of the temporal properties of language processing more generally. Spoken and signed languages make use of two different sensory channels (the acoustic and the visual channel); comparing the two offers a unique opportunity to investigate to what degree modality shapes the temporal structure of language. Our results suggest that temporal language processing arises from the interaction between the properties of the sensory system and the special characteristics of language. The perceptual modality poses constraints on how linguistic information is optimally processed by our cognitive system: the visual and auditory systems are characterized by different properties and perceptual processing of the physical signal will be different. At the same time, language structure accommodates the advantages and limits of a specific modality. This is particularly clear in the case of spoken and signed languages, where different temporal structures reflect the sequential or parallel organization of the information. Within the visual modality the results for the language signal show an advantage compared to non-linguistic material, suggesting the informational and temporal properties of the language signal favour its processing by our cognitive system. The reciprocal influence that language and sensory modality plays in shaping the temporal structure of the signal is extremely complex, and this study provides a first step towards disentangling their specific contributions. In Chapter 5 we further discuss the interplay between language structure and sensory modality, in light of the results from other studies presented in this doctoral work.

After investigating the role of temporal structure in language comprehension, in the following chapter we turn our attention to language production. We explore the physical characteristics of the visual linguistic signal in spoken and sign language: through the use of a motion tracking system we record three-dimensional motion data for two spoken and two signed languages, and apply different kinematic measures to describe and distinguish the two types of signals.

Chapter 3: Kinematic analysis of spoken and sign language visual signals

3.1 The importance for sign language research

The study of spoken language has devoted much research to how different acoustic properties of the speech signal are processed throughout the auditory system and mapped onto linguistic meaning (Poeppel & Assaneo, 2020). Since the invention of the sound spectrograph in 1945, acoustic analysis has been a powerful tool to characterize the physical properties of speech in terms of measures such as sound frequency, intensity, and duration. Conversely, the physical analysis of the visual signal in sign language is much less developed. This lag may be attributed to two main reasons: technology to record and measure the physical signal developed much faster (and is therefore much more available) for the acoustic domain than the visual domain, and the multi-articulatory nature of sign language adds a further layer of difficulty to this task.

Sign language makes use of several articulators to convey linguistic information: hands and fingers, head, arms, body and facial articulators (lips, gaze, and eyebrows). All these body parts are partly independent from each other, which allows them to move simultaneously and quasi-autonomously. Sign language production is characterized by a wide range of movements: from coarse movement such as torso oscillation during role-taking, to quite fine-grained movements performed by the fingers (Figure 6). These movements often include the hands touching each other or different body parts, creating visual occlusion of parts of the body. All these characteristics make the visual signal of sign language highly complex and, from a technical point of view, difficult to measure and quantify.

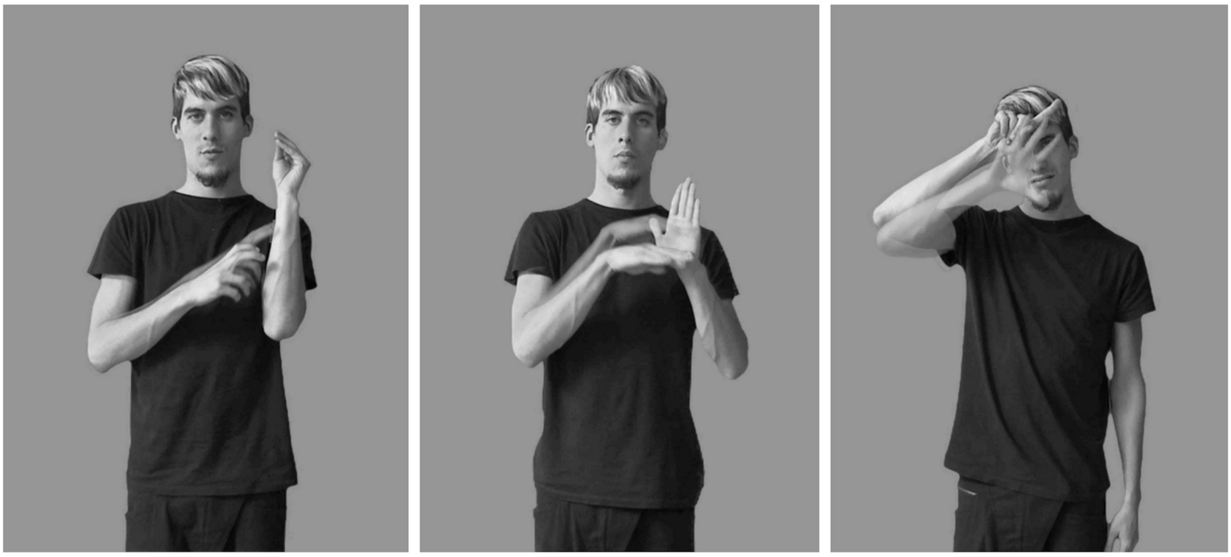

Figure 6:Example of signs showing different types of movement. From left to write: archaeology (arqueologia), programm (programa) and regret (arrepentimiento).

In order to characterize the physical properties of sign language production, we need a system capable of measuring location and movements of sign language articulators over time. Motion capture, commonly known as MOCAP (MOtion CAPture), refers to a heterogeneous group of techniques that makes it possible to track and record body movements over time and provides a possible tool to investigate language production in the visual domain. Traditionally, the tracking and recording is usually performed through a hardware device; more recently, motion tracking algorithms can detect features from videos recorded with standard video equipment. Depending on the technique employed, the output can provide different types of information describing the movement: either coordinates for each body part or a measure of the global visual change.

MOCAP technology has been used since the 70s mainly in professional applications in the cinema industry, but in the last forty years the increasing interest and the fast development of technology led to a multiplication of MOCAP systems and their availability for research purposes. The possibility of recording and analyzing high temporal and spatial resolution data of articulator movement during sign language production opened up many research opportunities, causing a rapid growth of this field in recent years.

Various research projects have attempted to characterize the properties of signed production, focusing mainly on cataloguing the different possible configuration of handshape, location and movement, and with the final aim of creating complete lexical databases describing sign properties (Caselli, Emmorey, & Cohen-Goldberg, 2021; Gutierrez-Sigut, Costello, Baus, & Carreiras, 2016). These studies are based on manual annotation of big datasets of videos, and are highly time consuming to perform. The advent of MOCAP technology simplified this task and opened the door to new lines of research. Point-light technology was one of the first motion tracking systems utilized to investigate sign comprehension with varying degrees of information (Poizner, Bellugi, & Lutes-Driscoll, 1981). This system, initially employed in biological motion research (Johansson, 1973), involves attaching small lights on the signer's body at strategic points (such as the head, shoulders, elbows, wrists and index finger tips), making it possible to record movements in a darkened room (Figure 7). These initial techniques required the person being recorded to wear some sort of hardware to record the movements. Other examples of model-worn hardware are data gloves with accelerometers measuring the magnitude and direction of hand movements. More recent work has instead exploited the potential of machine learning algorithms applied to single- or multi-camera recordings to identify the different body parts (Cooper, Holt, & Bowden, 2011).

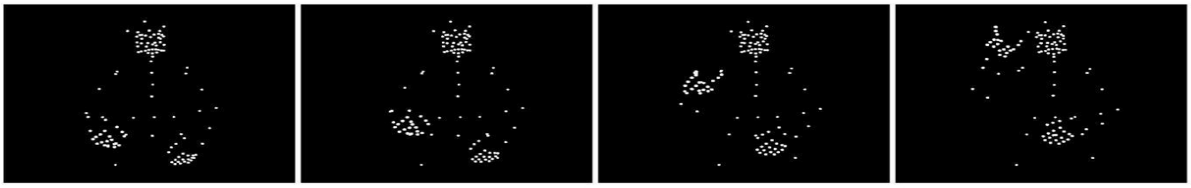

Figure 7: Example of a sign in ASL recorded and displayed with point-light technology (modified from Wikipedia. https://commons.wikimedia.org/wiki/File:Point_Light_Display_of_ASL_sentence.gif *).*

In the next section, we examine different types of MOCAP systems. Motion tracking technologies have been used in automatic translation from and to sign language. One application is sign language recognition, which aims to classify and identify signs based mainly on hand configuration to automatize sign-to-text and sign-to-speech translation (see Cheok, Omar, & Jaward, 2019 for a review of the literature). Another application of MOCAP technology is sign language synthesis, which aims at creating realistic avatars producing sign language (Elliott, Glauert, Kennaway, Marshall, & Safar, 2008). Its main applications are quasi-instantaneous translation from speech or text into sign language (Kahlon & Singh, 2021; San-Segundo et al.,

2008) and the creation of bilingual storybook apps to foster literacy development in deaf children (Malzkuhn & Herzig, 2013). The development of this technology ultimately aims at facilitating communication between signers and speakers. In the domain of research, recent motion tracking algorithms have been used to control, prepare and edit video stimuli in sign language experiments (Börstell, 2022; Trettenbrein & Zaccarella, 2021), and to characterize the temporal properties of the linguistic visual signal (Brookshire, Lu, Nusbaum, Goldin-Meadow, & Casasanto, 2017; Malaia, Krebs, Borneman, Wilbur, & Roehm, 2016).

3.2 Different motion tracking systems

The last twenty years have witnessed a proliferation of extremely different motion tracking systems. They vary in their characteristics and recording quality, in particular with respect to spatial and temporal resolution of the data. The accessibility of the methods and the technical expertise needed to operate them are also important varying factors, together with their cost (from a completely open access algorithm to a very expensive machine). Within this heterogeneous group of techniques, researchers have to find the MOCAP system which best fits their research needs. Adopting the classification used by Pouw and colleagues (Pouw, Trujillo, & Dixon, 2020), this section reviews tracking methods by dividing them in two major categories: video-based and device-based methods (Table 4).

Table 4: Overview of MOCAP methods

Method	***Type***	***MOCAP system***	***Spatial resolution***	***Temporal resolution***	***Dimensions***	***Cost***	***Technical skills***
Video based	Pixel differentiation	IVC	1 vector	Video frame rate	2D	Zero	Low
		Optical Flow					
	Machine learning algorithm	Open Pose	Body(15) +Face(70)	Video frame rate	2D/3D (with multiple simultaneous recordings)	Zero	Low - Medium
		Deeplabcut	Body(15) +Face(70)				
Device	Marker based		Body+Face	Video frame rate	3D	High	High
	Markerless	Kinect	Body(21)+Face(68)				

based	Leap Motion	Hands (26)	Variable	3D	Medium	High

The video-based category incorporates all those methods where a type of algorithm is applied directly to a standard video stream or recording. These methods have the advantage of being virtually costless as they do not require any special hardware, and they can be applied to any database of pre-recorded videos. The movement has to be inferred from two-dimensional data and the quality of the tracking depends on the quality of the video. One type of video-based method is *pixel differentiation*, where the magnitude of visual change across consecutive frames is calculated. Since this method relies on averaging some measure of visual difference across all pixels, it lacks any spatial resolution and cannot distinguish the movement of different articulators. Nevertheless, pixel differentiation methods are fairly easy and inexpensive to implement compared to specialized motion-tracking technology, and can provide a reliable estimation of gross-body movement (García-Bautista, Trujillo-Romero, & Caballero-Morales, 2017). To our knowledge two pixel differentiation methods have been used in sign language research: Instantaneous Visual Change (IVC) developed by Brookshire and colleagues (2017) and Optical Flow (Mcdonald et al., 2016). Another type of video-based motion tracking system is a *machine learning algorithm*, employed for both object and body detection. One of the most well-known implementations, widely used for research purposes, is the OpenPose algorithm. The OpenPose library (Cao, Hidalgo, Simon, Wei, & Sheikh, 2019) is a collection of real-time multi-person keypoint detection libraries for body, face and hand estimation. OpenPose is based on deep-learning algorithms: the RGB image is fed to a two-branch multi-stage convolutional neural network (CNN), which returns confidence maps and affinity fields of different body parts allowing for tracking over time. This method allows for 2D pose estimation with a single camera and 3D pose estimation with multiple cameras. Similar to OpenPose, Deeplabcut (Mathis et al., 2018) uses a deep neural network algorithm to detect the location of selected body parts. This system grants more flexibility in choosing which body features are subject to the tracking, and it is less automated in its implementation.

Device-based MOCAP systems require specialized hardware to perform motion capture, and usually have higher spatio-temporal resolution. Within this category some tracking systems employ markers (optic or electromagnetic) that are attached to the body of the person whose movements are being recorded. These *marker-based systems* may limit freedom of movement and the environment where the recording can be carried out. Due to the complex hardware they require, these systems are often quite expensive. In contrast, other devices use depth cameras that project infrared light on the scene and, based on the reflected light pattern, compute the distance of the different objects in the scene from the camera. These depth camera techniques make it possible to record the position of the body in a three-dimensional space without the need for any equipment on the subject's body. Although these *markerless systems* lose some temporal and spatial resolution with respect to marker-based systems, they allow for better ecological validity since the person can move naturally without having to wear any cumbersome equipment. In the case of sign language, this allows the model to sign more naturally, and the resulting recordings are more natural to watch since there is no extraneous material on the signer's body. Two well-established examples of markerless systems are LeapMotion and Kinect, which both rely on a combination of infrared cameras and motion tracking algorithms to track body movements in the three-dimensional space. The main downside of this type of system is that the tracking is possible only with special recordings from these devices, which means that recordings from other cameras cannot be tracked.

3.3 Dataset description

In the current study we recorded the motion tracking information of short videos of individuals telling stories in signed and spoken languages, so that these videos could serve as stimuli for the experiment on language-brain entrainment described in Chapter 4. Since we are interested in the specific kinematic profile of different articulators during language production we gave importance to the spatial resolution of motion tracking data.

To record the videos and simultaneously track body and face points we developed a custom-built system equipped with a Kinect v2 camera. Kinect is a motion sensing input device comprising two different cameras: a RGB camera recording the video information and an infrared camera, which measures depth information via time-of-flight calculation based on the time the emitted light takes to go from the camera to the object and back again. This device was initially developed by Microsoft as a gaming add-on for the Xbox console, but has been used and validated as a research tool for motion tracking (Otte et al., 2016; Trujillo, Vaitonyte, Simanova, & Özyürek, 2019). Kinect v2 uses machine learning (specifically, a random-forest algorithm) to identify the different body parts based on the tracking data recorded by its depth camera. This model tracks 21 body points and includes a module specifically for face tracking, both in two-dimensional and three-dimensional space. For our implementation, we used two software utilities – Vitruvius (Pterneas, 2017), a 3D motion tracking framework for Kinect, and Unity (Haas, 2014), an Integrated Development Environment primarily used to develop games and simulations – to develop a toolbox comprising two main modules. MOCAPrecorder allows users to record and save the video and motion tracking data in real time, while MOCAPeditor is designed to visualize the video and the motion tracking data and, if needed, to edit the tracked points in the three-dimensional space. The motion tracking dataset recorded with Kinect comprised 3D coordinates for 21 body points and 68 face points. Figure 8 show a visual representation of the tracking points both for body (A) and face (B); a complete list of the tracked points is provided in table A3 in Appendix 2. Although we collected motion tracking data for both face and body points, in this thesis we are presenting analysis limited to body points. Kinect output is characterized by a variable sampling rate, which depends on the computational requirements of the tracked movements and the computer performance. The recording was performed at BCBL, the Kinect v2 custom-built setup was used with an HP ProDesk 400 G7.

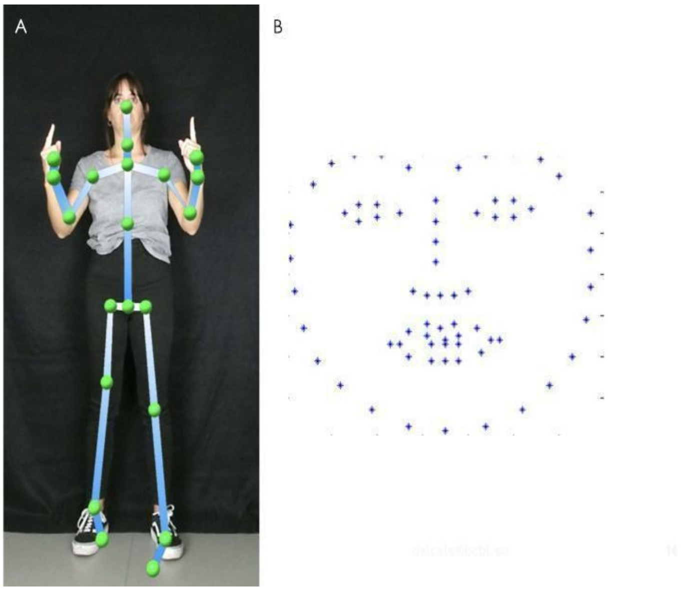

Figure 8: Display of the 21 body points (A) and 68 face points (B) recorded by our custom-built Kinect system.

Once the motion tracking system was operational, we recorded videos of semi-spontaneous speech and sign in four different languages: Spanish, Russian, Spanish Sign Language (LSE) and Russian Sign Language/Русский Жестовый Язык (RSL). For each language, we asked two native speaker/signer models to retell short narratives based on comic strips. Each comic strip featured recurrent characters interacting and engaging in different actions and events and did not contain any words or written language (see Figure 9 for example comic strips). We choose comic strips because visual narrative description is a widely used and robust method to elicit spontaneous speech and gesture (Cravotta, Grazia Busà, & Prieto, 2019). The same set of comic strips was used to generate the narratives in all four languages. Eight models recorded 50 videos describing 50 different comic strips, out of this sample we selected 40 videos in each language to be used for subsequent analysis and for the experiment described in Chapter 4. For each language we had a male and a female model, with the exception of Russian Sign Language (for which both models were male due to the limitations of finding native signers); all models were right handed. Each video was approximately one minute in length and filmed against a uniform black background with controlled lighting. Models familiarized themselves with the comic and then retold the story as if they were telling it to a friend. Models started and ended each video with their hands in resting position, and could self-regulate using a timer positioned above the camera. They were instructed to remove or add details to the story as they pleased, in order to make the speech and sign as natural as possible. The videos had an average length of 60.59 s (SD= 3.91 s). The recording was performed at BCBL with the Kinect v2 custom-built set up described above, which allowed the recording of both video and motion tracking information for each narrative; in the case of spoken languages audio was recorded as well.

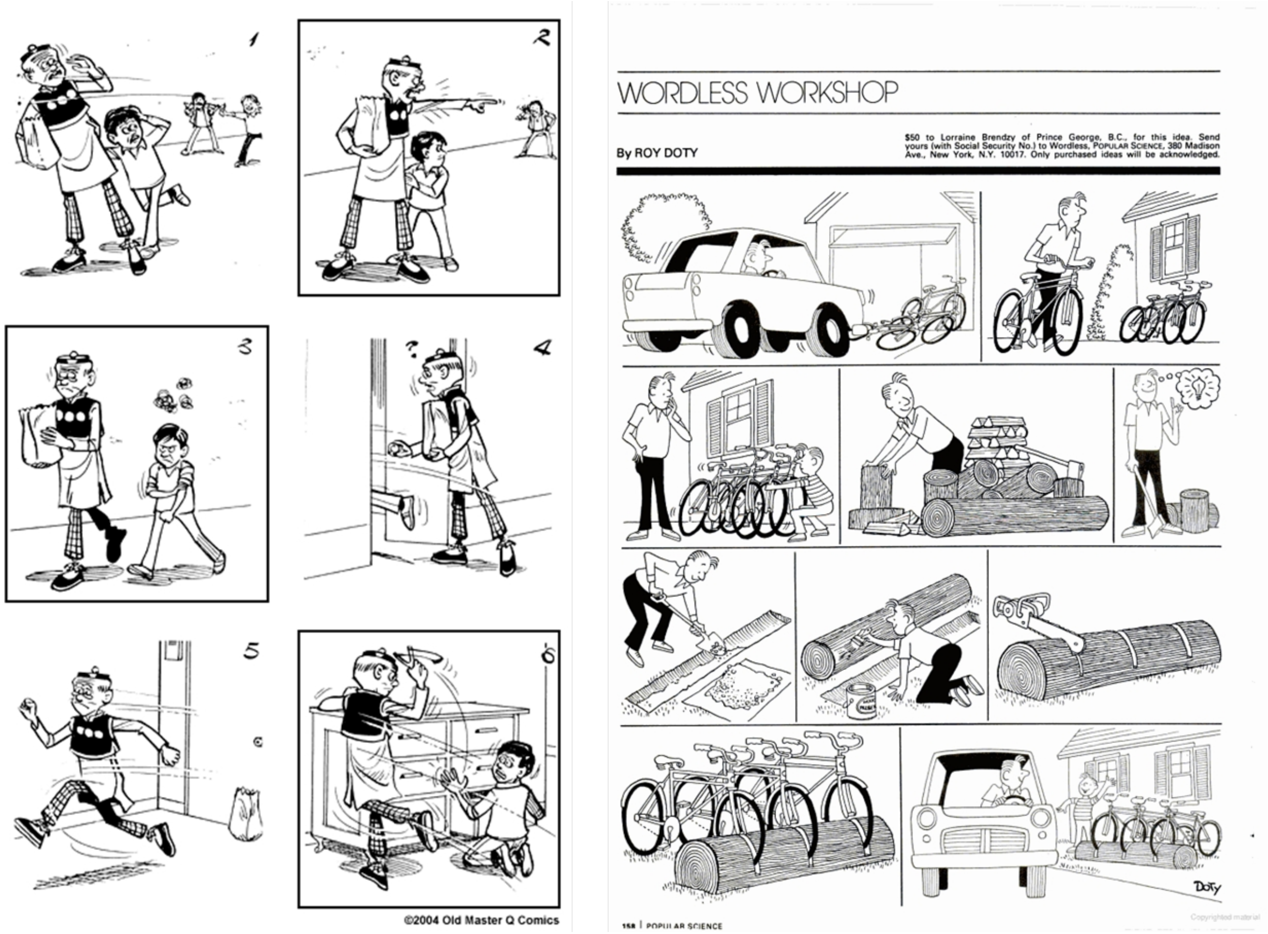
1
2
3
4
5
6
©2004 Old Master Q Comics
WORDLESS WORKSHOP
By ROY DOTY
$50 to Lorraine Brendzy of Prince George, B.C., for this idea. Send yours (with Social Security No.) to Wordless, POPULAR SCIENCE, 380 Madison Ave., New York, N.Y. 10017. Only purchased ideas will be acknowledged.
DOTY
158 | POPULAR SCIENCE

Figure 9: Example of two comic strips used to generate the videos.

3.4 Kinematic properties of language

The motion tracking data recorded with the Kinect setup is a rich dataset of 3D coordinates tracking the movement of several body parts over time, and gives us the opportunity to explore the kinematic properties of the visual signal of sign and spoken language. Even though motion tracking techniques have been used with sign language data for a variety of purposes (see section 3.1 for an overview of these applications), to date no consistent descriptive measures of the sign language signal have been identified and studied. In this section we draw from the literature on speech signal analysis and motion tracking for gestures, and adapt some of these analytical techniques to our dataset. Our aim is to identify different kinematic measures that are useful to characterize and describe the sign language (LSE and RSL) and spoken language (Russian and Spanish) visual signals, and that are possibly able to distinguish between the signals from the two language modalities.

3.4.1 Analysis

After data were preprocessed (as described in section 3.4.1.1), we conducted three different types of analysis: kinematic features (section 3.4.1.2.), time-frequency analysis and Uniform Manifold Approximation and Projection (UMAP) analysis (section 3.4.1.3.). Following the literature on motion tracking of gestures, we extracted kinematic features to identify those properties deriving from body movements that can be relevant to describe the linguistic visual signal: we focus mainly on movement segmentation, space and magnitude. Time-frequency analysis shows the periodicity in the signal and the specific frequencies at which this periodicity emerges. In spoken language studies this measure has proven to be relevant for speech processing (Poeppel & Assaneo, 2020). We are also interested in assessing whether time-frequency patterns cluster differently depending on different models and languages. We used UMAP, a dimension reduction technique that uses graph layout algorithms to arrange high-dimensional data in a low-dimensional space while preserving as much as possible the structure of the original dataset (McInnes, Healy, & Melville, 2018).

All the analyses were performed in R Studio, version 4.0.3 (R Studio Team, 2020). All scripts were written in RMarkdown and are available on the following public Open Science Foundation repository: https://osf.io/cmrav/?view_only=c9486cfe58d443a3ae8afefd91272078.

3.4.1.1 Preprocessing

A visual inspection of the tracking quality of Kinect recordings was performed through the MOCAPeditor toolbox. This inspection revealed the presence of small inaccuracies which can be ascribed to poor tracking performance and occlusions of certain body parts during movement. A cleaning algorithm was developed to fix and smooth these inaccuracies and improve the quality of the motion tracking (Pastureau, 2022). This algorithm measures movement speed over a certain amount of frames and through interpolation corrects coordinates for biologically impossible movements: when movement over three frames was 10 cm or more it was considered inaccurate and therefore corrected. This first preprocessing step was run on raw output data from Kinect.

Motion tracking and RGB data recorded with Kinect have a variable frame rate (mean = 21.64 Hz, SD = 2.62 Hz, min = 1.41 Hz, max = 31.37 Hz), Table A4 in Appendix 2 provides descriptive statistics of frame rate for each language. Both motion tracking and RGB data were resampled at 25 Hz before any subsequent analysis. Out of all body points tracked we focused on five body parts that represent linguistically informative articulators in sign language: head, torso (in Kinect labeled as SpineMid), right hand, left hand, right shoulder. All subsequent analyses were run on the motion tracking data collected for the selected articulators. We applied a first-order 10 Hz low-pass Butterworth filter to smooth out high-frequency jitters usually caused by motion artifacts present in Kinect raw data. To disentangle the absolute motion components of each articulator from the overall body movement we re-referenced the coordinates by subtracting the motion of the torso from that of all the other articulators.

3.4.1.2 Kinematic features

Based on Pouw and colleagues (Pouw et al., 2021) we selected three features to describe the kinematic profile of our data: submovements, rhythmicity and motion space.

1) The *submovements* measure (Trujillo et al., 2019) is based on the velocity of the articulator and is computed with a peak finding function to identify and count the peaks in the speed time series of each articulator. Based on the minimum time interval between consecutive peaks (set at 1 second) and the minimum peak height (in our case calculated as one standard deviation below the mean peak velocity), this function isolates specific movements based on their acceleration and deceleration profile. This measure provides a useful (although less reliable) alternative to manual coding of individual movements. Figure 10 shows the right hand speed time series for one video in LSE, and the 37 submovements identified by the peak-finding function. Some peaks, although higher than the threshold peak height value (shown as a dotted line in the figure), are not considered submovements because they are too close to other peaks and therefore do not meet the requisites of the minimum peak distance.
2) *Rhythmicity* reflects the temporal variability of the extracted submovements and is calculated as the standard deviation of the time intervals between consecutive submovements. A higher rhythmicity value reflects more temporal variability in the

movement; in contrast, a lower score is indicative of a more regular rhythm. Figure 10, for example, shows a rhythmicity value of 0.01 indicating a highly isochronous rhythm.

3) *Motion space* indexes the size of the spatial envelope that each articulator moves in and is calculated as the square root product of the differences between the minimum and maximum coordinate points in each dimension (x, y and z). This measure indicates how much space the articulator movement takes up. We also calculate *motion magnitude* – the average of each articulator's speed vector – to describe how much the articulator actually moves during the entire recording (independently of the size of the spatial envelope). Thus a single large movement would have a large motion space and a small motion magnitude (i.e., it takes up a lot of space but has not moved a great deal); in contrast, a small circular movement that is repeated many times would have a small motion space, but a large motion magnitude (i.e., it takes up little space but has moved a lot).

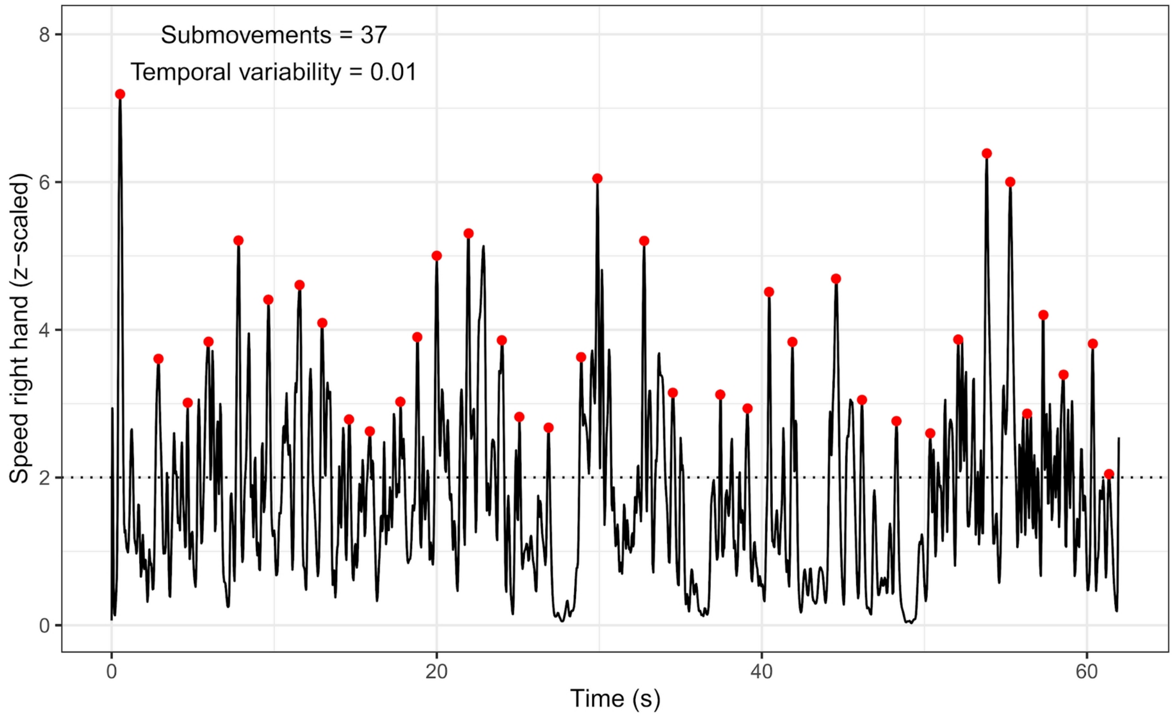

Figure 10: Speed vector of one video in LSE showing right-hand movement. Submovements measure is shown as the peaks in red.

The speed vector of each video was calculated on pre-processed coordinates in the three-dimensional space, and the resulting vectors were smoothed again with a Butterworth filter. To compare the kinematic patterns of each articulator among different languages, we independently normalized the speed vectors for each articulator across all videos (including all models and all languages). This type of normalization allows us to compare, for example, the number of submovements of the right hand between LSE and Spanish. Conversely, it prevents comparing one articulator with another. *Submovements*, *rhymicity* and *motion magnitude* were extracted from the normalized speed vectors. *Motion space* instead was calculated on pre-processed coordinate data which did not undergo normalization. Results for sign language and the visual speech signal are shown in the section 4.2.1 below.

3.4.1.3 Time-frequency analysis

First the speed vector of each video was calculated on pre-processed coordinates in the three-dimensional space, and the resulting vectors were smoothed again with a Butterworth filter. Fast Fourier transformation analysis was calculated on these speed vectors and the resulting power data were averaged in frequency bins of 0.2 Hz from 0 to 12 Hz. The power vectors were then normalized with a similar procedure described above: power vectors were scaled across all videos separately for each articulator. UMAP analysis was applied to assess whether time-frequency patterns cluster differently depending on different models and languages. Plots for the time-frequency analysis and the UMAP clusters are presented in section 3.4.2.2.

3.5 Results

Below we present the plots describing kinematic features, time-frequency analysis and UMAP representation for all articulators under study.

3.5.1 Kinematic features

Boxplots in Figures 11 and 12 show the results for *submovements* and *rhythmicity*: we compare the four languages (Spanish, Russian, LSE and RSL) for each articulator by averaging across models and videos in each language. Torso submovements are minimal in all languages, although LSE shows greater activity compared to the other three languages. In contrast, the other articulators show greater activity, and differences between the different languages and modalities. Head, right hand, left hand and right shoulder are all characterized by a higher number of submovements in signed languages compared to spoken languages. For the spoken languages, there is virtually no activity of the head and right shoulder, while Spanish has more hand movements than Russian does. For the sign languages, there is little difference between LSE and RSL, with a comparable number of movements for each articulator.

The results for rhythmicity do not highlight specific differences among languages, although overall there is a trend of sign languages being slightly more regular in rhythmicity (i.e., having lower scores) compared to spoken languages. Rhythmicity for the torso could not be calculated in RSL, spoken Spanish and spoken Russian due to a lack of submovements.

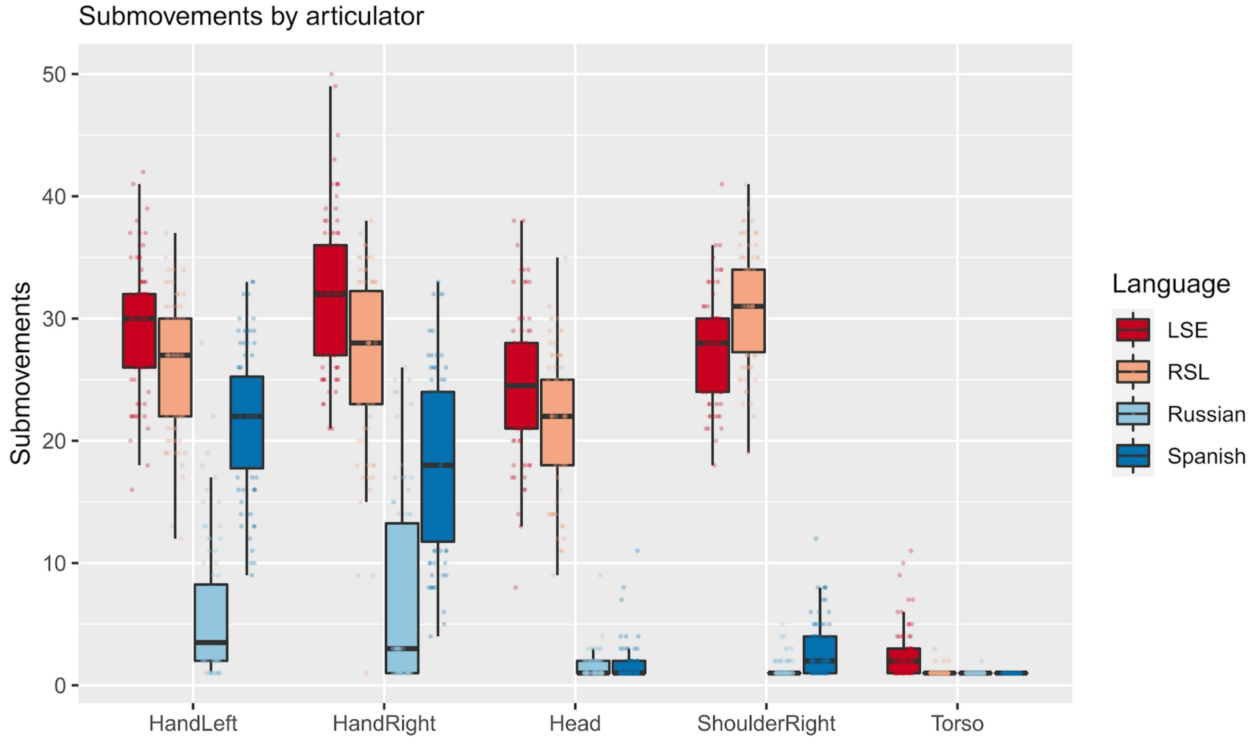

Figure 11: Number of submovements for each articulator across the four languages.

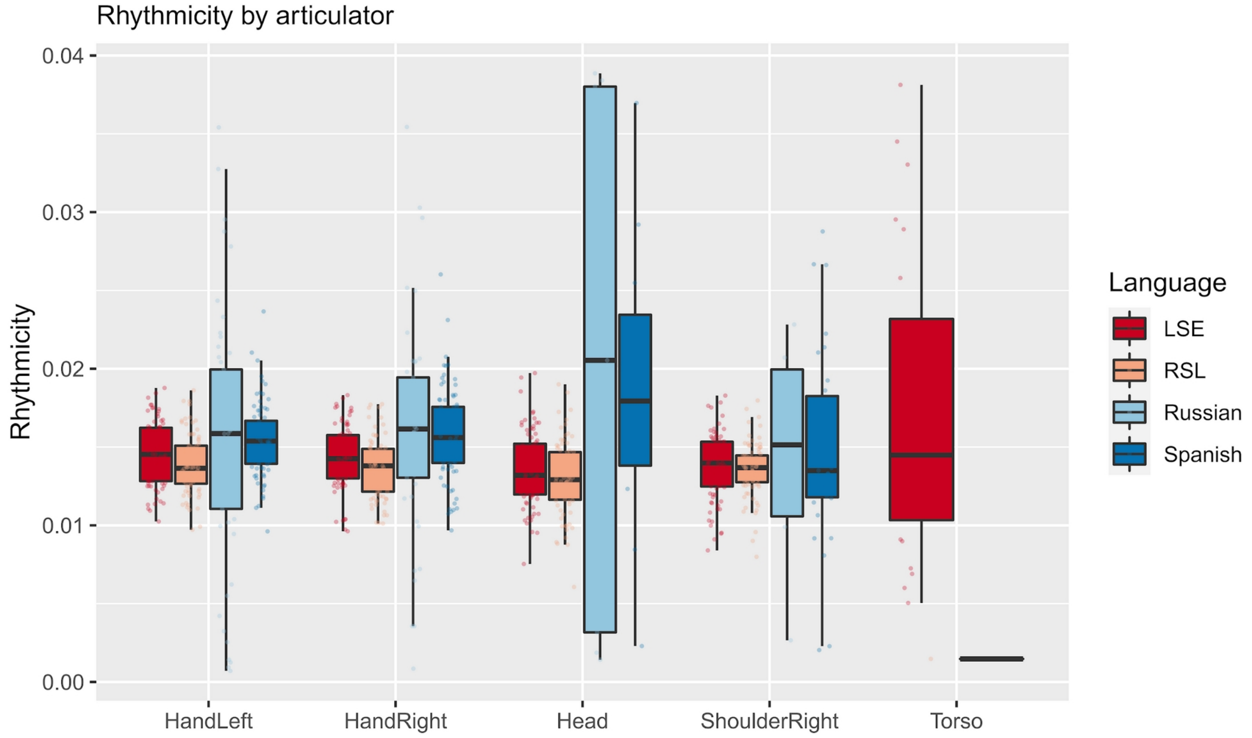

Figure 12: Rhythmicity for each articulator across the four languages.

Motion space and motion magnitude measures are presented in Figure 13, where the points on each silhouette present the average values across all videos and both models of each language. Motion space for each articulator of interest is indexed by the size of the dots, while motion magnitude is indexed by the colour of the dots. Note that the motion space is based on the raw (unscaled) data, which means that it is possible to compare this measure across articulators. Often the two measures go hand in hand: an articulator with a large spatial envelope (i.e. motion space) tends to have moved more (i.e. motion magnitude). The torso, head and right shoulder show similar motion space and magnitude across the four languages. Conversely, both right and left hands show greater use of space and more overall movement in signed compared to spoken languages.

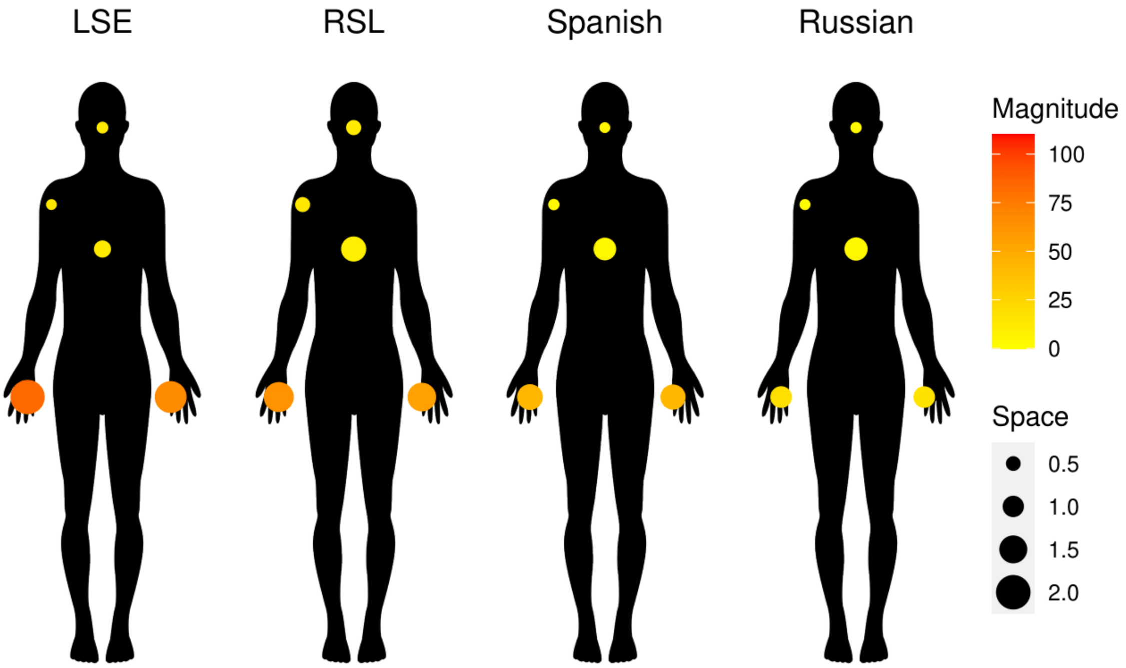

Figure 13: Motion space and motion magnitude for each articulator across all four languages. Motion space measure is indexed by the size of the dots representing the articulators, motion magnitude is indexed by the colour.

Signed languages show a high degree of overlap, pointing to a similarity in kinematic features in this modality independently of the specific language. Compared to the sign languages, the spoken languages do not show as much uniformity, a result which is not surprising considering that co-speech gestures show a high variability across languages and speakers (Kita, 2009).

3.5.2 Time-frequency analysis

Figure 14 shows the overall time-frequency pattern for each model in each language, for all articulators combined and also for each articulator individually. The power spectrum for each model was extracted averaging power values across all 40 videos recorded by one model. Power spectrum plots averaged for each language are available in Figure A1 in Appendix 2. The time-frequency pattern of the combined articulators shows a clear increase of power in all frequencies for signed languages compared to spoken languages. This difference reflects two main characteristics of sign language: the increased amount of movement in sign language (as previously shown by submovements and, to some extent, motion magnitude measures) and the higher level of ‘periodicity’ of this movement. Looking at each articulator’s specific time-frequency profile we can see that the same pattern is present for right and left hands, right shoulder and head. The torso does not show any clear difference in power across different models or languages.

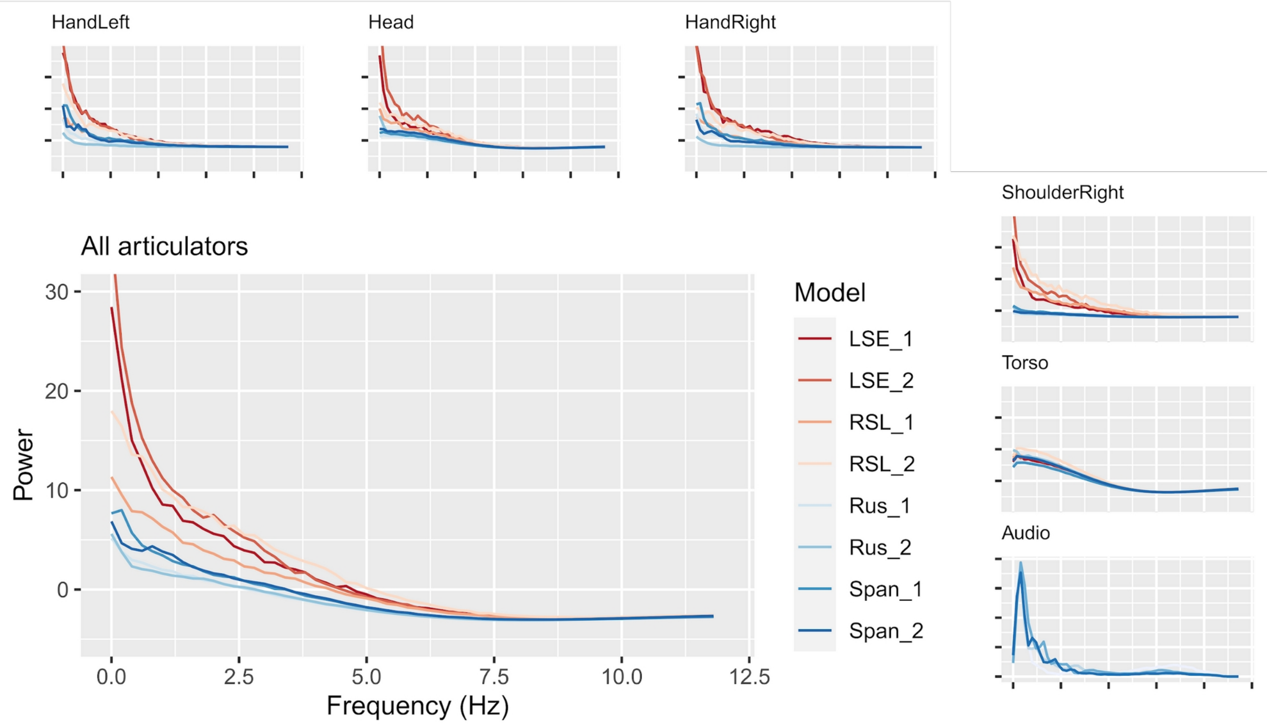

Figure 14: Time-frequency plots across models and languages. The central plot represents the power spectrum of all the articulators of interest, with smaller plots for each articulator. The power spectrum for speech in Spanish and Russian is shown in the bottom right corner.

UMAP analysis was used to investigate which videos are more kinematically similar to each other based on their time-frequency properties. Figure 15 shows the time-frequency data grouped by language for all articulators of interest in the large plot, and for individual articulators in the smaller plots. In the plots, each point represents one video from our dataset, while each colour maps onto a specific language. UMAP plots clustered by models are available in Figure A2 in Appendix 2. Interestingly, when taking into account the time-frequency patterns of all articulators of interest UMAP succeeds in creating clear clusters of videos for each language. Moreover, the two signed languages and the two spoken languages are closely clustered suggesting that languages in the same modality are kinematically more similar to each other than to languages in a different modality. Similarly, time-frequency patterns for right shoulder and head show clusters that distinguish language modalities. This is less so for the hands, especially the left hand, where clusters from different modalities overlap. Torso, labeled as spine-mid, fails to highlight any discernible difference between languages or modalities.

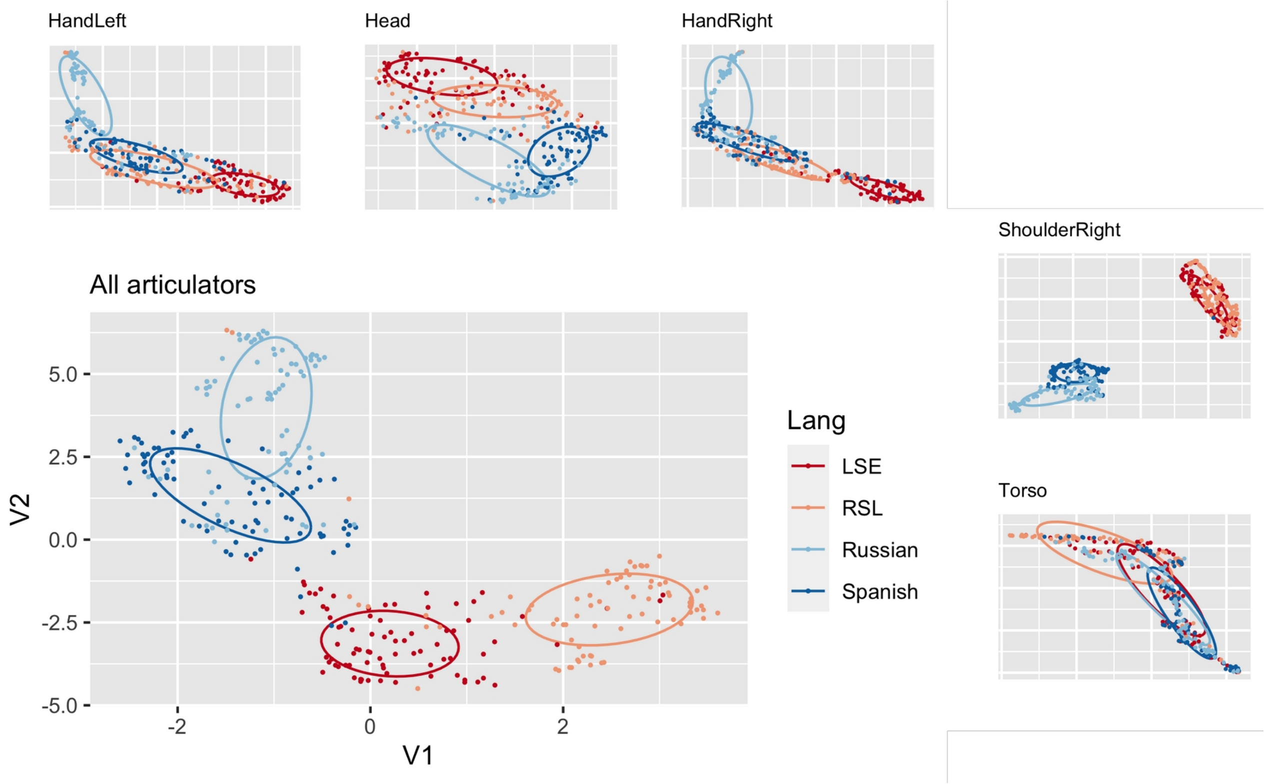

Figure 15: UMAP clusters across the four languages. Each dot represents a video of our dataset, and the classification is done based on the time-frequency profile derived from each video. The central plot shows the clusters based on all the articulators of interest, with smaller plots for each articulator.

3.6 Discussion

Due to its intrinsic complexity the study of motion and visual characteristics of sign language is still predominantly based on qualitative description and analysis. The use of motion tracking techniques to record the movements that occur during sign language production, as well as those that accompany speech production, opens up various possibilities to characterize the visual properties and kinematic features of language.

In this chapter we evaluated the suitability of four different kinematic features (*motion space*, *motion magnitude*, *submovements* and *rhythmicity*) in describing and characterizing the visual signal present in spoken and signed language production. Motion space, motion magnitude and submovements proved to be informative: as expected, sign languages are characterized by more submovements as well as a larger use of space and more movement by the hands. Rhythmicity, on the other hand, failed to highlight any meaningful difference between the two language modalities. This measure might not be sensitive enough to pick up periodicity in the signal (as shown instead in the time-frequency plots), but this lack of differences can also be due to averaging data across all sentences and models within a language. Another option is that the use of quasi-naturalistic stimuli makes our data too 'noisy' to show any difference between language modalities in this measure. The analysis of the time-frequency patterns of movements in spoken and signed languages also reveals a clear difference between the two modalities: sign language is characterized by higher power in all frequency bands, in line with the idea that during sign language more periodic movements are produced compared to spoken language. This difference is confirmed by the results of UMAP analysis, which reliably distinguishes between spoken and sign language time-frequency patterns and, in certain cases, can even isolate specific languages.

Our analysis, although explorative, shows the validity of kinematic analysis to characterize the different types of body movement during language production. These results suggest that, even if gestures and signs are produced via the same articulators, their temporal structures are quite different. Gestures are produced simultaneously with speech and in this setting these two signals (visual and acoustic) tend to couple (see Wagner, Malisz, & Kopp, 2014 for a review on the topic). Prosodic contrasts in speech, for example, structurally align with peak velocity of gestures (Danner, Barbosa, & Goldstein, 2018; Pouw & Dixon, 2019). In multimodal spoken language the acoustic and the visual signals, which belong to two modalities characterized by very different temporal structures, need to interact and combine to create a meaningful linguistic output. In sign language instead, the visual signal is the only channel used to deliver linguistic information and it is interesting to notice that, in the absence of the acoustic modality, body movements tend to temporally organize in a different way.

Overall, our analysis highlights the role that different articulators play in language depending on the modality. Right hand and shoulder reliably differ between spoken and sign language in all the kinematic features investigated. Given that all the signing models were right-handed, this difference can be ascribed to the important role that the dominant hand plays in sign language production: one-handed signs and fingerspelling are always produced with the dominant hand and in many two-handed signs the non-dominant (in this case, left) hand plays a more supportive role. The different kinematic profile of head movements between spoken and sign languages is captured by submovements and time-frequency patterns, but not in the motion space and magnitude measures, suggesting that it is not how much the head moves but rather how it moves that distinguishes speech from sign. Turns and tilts of the head are linguistically relevant movements in sign language, but the motion tracking data analyzed here are not capturing this specific type of motion. Motion tracking of face points would make it possible to distinguish these different types of movements and thus might be critical to capture these differences between spoken and sign language. Finally, the torso does not represent a decisive articulator in differentiating sign and spoken language movements; this suggests that translational movements of the body are not relevant in sign language production and fits with the observation that signers normally stand stationary while signing, just as speakers do. These findings support the importance of employing motion tracking techniques with fine spatial resolution, and therefore being able to disentangle the specific kinematic properties of different body parts.

Kinematic features also seem to be much more similar between sign languages than between spoken languages. Research has shown that several cultural factors, as well as the structure of the spoken language, influence co-speech gesture production (see Kita, 2009 for the cross-cultural variation of gestures). Moreover, the use of gesturing is not obligatory in spoken language communication and shows great variation depending on usage situations and individuals (Kita & Özyürek, 2003; Mcneill & Quek, 2005; Streeck, 2009). Conversely, visible body movements in sign languages are the primary signal and information bearers, which means that there should be less heterogeneity (although with some level of variation) in the kinematic properties of sign language.

The motion tracking analysis presented in this chapter is preliminary and suffers from some limitations. The sample of signers is very small; to strengthen and generalize the results, data from several signers and several sign languages need to be collected and analyzed. Our findings seem to point to some universal kinematic properties shared by sign languages, but more research is needed to characterize possible differences across different signed languages and signers (taking into account individual variables such as language proficiency, hearing status and age of acquisition). Although our analysis was able to reliably show specific differences between spoken and sign language based on the articulator, we are aware that facial articulators need to be included for a complete account of sign language kinematics. Finally, an important clarification needs to be made regarding our dataset: number of signs and gestures in the videos are not matched and the unbalanced presence of movements might represent an element of confound in the results. Nevertheless, the videos analyzed in this chapter were recorded with the purpose of eliciting naturalistic speech and sign from the native models, therefore we expect them to be a realistic representation of the visual signal that available during speech and sign produced during real-life language usage. While we are aware that our dataset does not offer a balanced comparison between sign and co-speech gesture, we believe that these findings can still represent a valuable contribution to the understanding of motion kinematics of language.

This field of research is still in its embryonic stage, but it already shows a fast growth thanks to the development of accessible motion tracking techniques and promising results. Research focusing on multimodal language, in particular co-speech gestures, is already making extensive use of these methods and the sign language field can adopt the analysis techniques developed for kinematic gesture data and adapt them to the study of sign kinematics. The step is not trivial, as gestures are usually studied in isolation while sign is characterized by a more continuous stream of movement. Future research in this field will need to investigate which different measures capture relevant kinematic properties of the linguistic visual signal, and eventually link these features with linguistic properties of the signal based on models of sign language phonology and syntax. The preliminary analysis presented in this chapter clearly shows the promising value of motion tracking analysis applied to the linguistic visual signal, in particular in the domain of sign language. Characterizing the visual properties of sign language will help better understand the complex relationship between production, perception and brain processing of language; with the aim of filling the gap between spoken and sign language literature. This issue is explored in the next chapter.

Chapter 4: Language processing in the brain

4.1 Introduction

The study presented in the current chapter investigates whether language-brain entrainment is a phenomenon restricted to spoken language, or whether it extends to a language expressed and perceived through the visual modality such as sign language. We took into account two main factors: language modality, by including spoken and signed languages, and language knowledge, using languages that were known or unknown to our participants. Neurophysiological activity of two groups of hearing participants, proficient signers and sign-naive individuals, was recorded with MEG while they were watching videos of storytelling in Spanish Sign Language, Russian Sign Language, spoken Spanish and spoken Russian. The design of the study allows us to disentangle specific effects of modality and linguistic experience, as well as familiarity with a language modality, and thus to better understand how sign languages are processed and, more generally, the role that entrainment plays in language processing.

4.1.1 Effect of language familiarity on entrainment

As mentioned in section 1.4, language-brain entrainment is a well documented phenomenon involving the temporal alignment between brain activity and a perceived speech signal. This temporal alignment specifically consists of the synchronization of neural activity with the rhythmic properties of the speech envelope (Obleser & Kayser, 2019), and may be driven by both bottom-up and top-down processing, although the relative contribution of each is still unclear. In the first case, entrainment is elicited solely by the exogenous stimulus: the periodicity of the stimulus' temporal structure (or its presentation rate) drives entrainment. Top-down processes such as attention or linguistic knowledge also play an important role. Paying attention to a specific speech stream in a noisy environment has been used to investigate attentional top-down effects on entrainment (Mesgarani & Chang, 2012; O'Sullivan et al., 2015). Such studies demonstrate that cortical tracking of speech is not equally elicited by any speech stream in the environment, but is modulated by direct attention. Other studies have tested the top-down effect of language familiarity on entrainment, with mixed results. These studies manipulated familiarity by testing groups of participants with different levels of language proficiency. When presented with speech in Chinese, native speakers of American English failed to entrain to the periodic syntactic structures (such as phrases and sentences) in the stimuli. Periodicity in syntactic structures which are not marked by auditory cues in the exogenous signal is tracked only when participants know the language (Ding et al., 2016). Similarly, language proficiency may modulate cortical tracking of regularities in speech in theta and delta (Lizarazu et al., 2021) or gamma frequency bands (Peña & Melloni, 2012). Finally, Brookshire and colleagues (2017) showed that entrainment to sign language changes in strength and topographical distributions depending on language knowledge.

The literature investigating the effect of language familiarity on entrainment converges in showing that to a certain degree entrainment is automatically elicited by any quasi-periodic signal (including unknown languages), but is modulated by top-down processes such as attention or language knowledge. The nature of this modulation is not completely clear: some studies have shown differences in the strength of the entrainment (i.e., power of coherence) in different frequency bands, while others reveal changes in the topography of the effect (Brookshire et al., 2017; Ding et al., 2016; Lizarazu et al., 2021; Peña & Melloni, 2012). These diverse findings are possibly due to the different types of stimuli employed and the different levels of proficiency of the participants.

The top-down effect of language knowledge interacts with the temporal structure of the language input itself. Languages are usually classified into three different categories based on their rhythmic structure (Abercrombie, 2019; Nespor et al., 2011): syllable-timed languages (e.g., Spanish, Italian), stress-timed languages (e.g., English, Russian) and mora-timed languages (e.g., Japanese). These different types of rhythm reflect different speech segmentation units (Cutler et al., 1986; Mersad et al., 2011; Ramus et al., 2000), and indeed infants seem to be able to discriminate between syllable and stress-timed languages from a very young age (Nacar Garcia et al., 2018; Nazzi et al., 1998). The rhythmicity of syllable-timed languages is driven by syllable rate, which overlaps with theta frequency bands (4-7 Hz); in stress-timed languages, rhythmicity is linked with stressed syllables and therefore should fall into lower frequency bands, such as delta (0.5-2.5 Hz). The effect of these different temporal structures on language-brain entrainment has not been studied yet.

4.1.2 Entrainment in the visual modality

Entrainment is not limited to the acoustic domain; in fact, it seems to be a more general mechanism that our cognitive system employs to optimally process different stimuli with quasi-periodic characteristics (Lakatos et al., 2019). This is also true for stimuli presented in the visual domain: rhythmic patterns in the activity of the perceptual system supports sampling of the visual input stream. One clear example is the entrainment between neural oscillations and eye saccades at around 3-5 Hz (Bartlett et al., 2011; Hoffman et al., 2013; Ito et al., 2011), which subserves the parsing of visual information.

Even if language-entrainment studies have focused for the most part on the acoustic speech signal, spoken language is multimodal in nature and uses the visual channel to convey facial expressions and gestures (Özyürek, 2014). A few studies have investigated entrainment during audiovisual speech, focusing on the role of mouth movements in speech comprehension. Results show entrainment to mouth movements at frequencies between 1–8 Hz in the primary visual cortices, reflecting the extraction of visual features (Bourguignon et al., 2020; Crosse et al., 2015; Park et al., 2016). Taking a different direction, a couple of studies looked at language in the visual domain employing sign language. Brookshire and colleagues (2017) found coherence in the frequency range 0.5-5 Hz between sign language input (ASL presented as videos) and brain oscillations recorded with EEG (see section 1.4 for a detailed description of the study). The study included both participants who did and did not know sign language. The synchronization between EEG oscillations and frequencies in the ASL visual signal was the same for signers and non-signers in occipital areas corresponding to the primary visual cortex, suggesting that both groups of participants pick up on perceptual regularities in the visual signal. ASL signers showed more entrainment in frontal areas compared to sign-naive participants, pointing towards an effect of language knowledge modulation. In a recent study, Malaia and colleagues (2021) compared the visual change in sign language videos with the EEG recordings of expert signers while watching those stimuli. A machine learning approach was applied to assess the contribution of the different frequency bins in predicting cortical coherence, and therefore language comprehension. The model yields highest prediction accuracy in lower frequencies up to 4 Hz, while higher frequencies (4-12.5 Hz) appeared to contribute less to explaining language-brain entrainment.

The few studies investigating language-brain entrainment in the visual domain all suggest that the auditory and visual channel share a similar mechanism: neural oscillation synchronises with the periodic component of the presented stimuli (whether visual or auditory). Sign language, by virtue of its complex linguistic structure and the exclusive use of the visual modality, represents the perfect test case to investigate language-brain entrainment. The main focus of this study is to identify whether we can find brain-language entrainment with a visual, signed language. To do so, we will use motion tracking data as a means to measure periodicity in the visual signal (see Chapter 3 for a detailed description of the motion tracking procedure), and relate it with the oscillations in the brain signal.

4.1.3 The experiment

The study presented here examines the functional role of the entrainment mechanism for language processing and understanding. On one hand, multiple studies provide evidence that entrainment plays a facilitatory role in linguistic processing: cortical entrainment with the speech envelope correlates with language intelligibility (Doelling et al., 2014; Gross et al., 2013; Peelle et al., 2013) and language proficiency (Lizarazu et al., 2021). Moreover, language deficits co-occur with a decrease of entrainment, as observed in poor readers (Abrams et al., 2009) and dyslexic children (Cutini et al., 2016). This evidence has led to the claim that entrainment tracks meaningful linguistic units within the linguistic sensory input, supporting its parsing and processing (Ghitza, 2011, 2013; Ghitza & Greenberg, 2009; Poeppel, 2003). According to another school of thought, entrainment is a sensory processing mechanism that applies to any type of signal independently of its linguistic content. This view is corroborated by studies showing entrainment in several domains outside of language (Lakatos et al., 2019) and the observation that the some of the frequencies linked with linguistic units are present in the auditory cortex at rest (Giraud et al., 2007). Cummins (2012) advises caution in identifying the syllable as the periodic unit at the base of language-brain entrainment, and claims that speech does not show enough periodic properties to scaffold cortical entrainment.

Here we focus on two specific questions in this extensive debate: the effect that language modality and language knowledge have on entrainment. Firstly, is entrainment in language merely due to the temporally evolving auditory signal that is speech, or does language entrainment occur in other modalities? We modulate the modality of the linguistic signal by comparing spoken and signed languages. The majority of the studies on language-brain entrainment employ spoken language stimuli exclusively in the acoustic modality. In this study we account for the visual component of language by presenting audiovisual spoken language material. Sign language is produced and perceived solely through the visual modality, and therefore represents a suitable case to test whether visually presented linguistic information gives rise to cortical tracking similar to what has been found for speech. Secondly, is language entrainment driven solely by the perceptual properties of the input signal, or does familiarity with the language impact how the brain syncronises with that input? We included a known and unknown language in each modality (spoken and sign) to evaluate the effect that language knowledge has on entrainment and investigate how it interacts with language modality. Alongside a group of hearing bimodal bilinguals, with a high level of proficiency in both Spanish and LSE, we tested a control group of matched hearing participants with no knowledge of sign language. The use of two groups allowed us to vary the level of experience in language processing within the visual modality, as learning a visual language has been associated with improvement in allocation of attention during visual processing (Bavelier et al., 2000; Proksch & Bavelier, 2002).

Based on the design of this study our hypotheses are as follows. In the known spoken language condition we expect to reproduce, with quasi-naturalistic stimuli, the typical results found in language-brain entrainment: marked entrainment in delta and theta frequency bands. We do not expect any difference between bimodal bilinguals and controls, since participants in both groups are native speakers of Spanish and have no knowledge of Russian If entrainment is causally implicated in language processing in a broader sense, then we expect to find this phenomenon for both spoken *and* signed languages. Conversely, if entrainment is elicited solely by processing in the acoustic modality, we expect no cortical tracking driven by sign language. Our expectation is that we will find entrainment for the sign language condition. Given the different temporal structures that characterize spoken and sign language, we also expect any entrainment to occur in different frequency ranges across modalities; for sign language we expect entrainment at lower frequencies, particularly in the delta band. The comparison of coherence between known and unknown languages within each modality will allow us to investigate to what degree language knowledge modulates entrainment, and to test the theory that our cognitive system employs entrainment for language comprehension. Based on the existing literature, we predict that entrainment to periodicity that forms part of the physical signal will be distributed mainly in primary sensory cortices (auditory and visual) for both known and unknown languages. Known languages should additionally show stronger cortical tracking in these areas and possibly also in frontal areas associated with higher level language processing. The comparison of entrainment to sign language between expert signers and non-signers will allow us to assess to what extent entrainment is driven by merely seeing a dynamic visual signal and how much is due to actually processing the signal as linguistic input. We expect bimodal bilinguals to show higher entrainment to sign language compared to controls. The design of this study lends itself to examine the impact of familiarity with a visual language on visual processing: we can compare the two groups in different conditions to see whether prior experience with a sign languages impacts how participants entrain to visual input. We predict that knowing a sign language could facilitate greater entrainment not only to a known sign language but also to an unknown sign language and possibly to the visual components of spoken language, compared to that of the sign-naïve participants.

4.2 Methods

4.2.1 Participants

A total of 33 participants took part in the experiment, divided into two groups. The first group was composed of 16 bimodal bilinguals who were native speakers of Spanish and native or highly proficient users of LSE. Six participants were native signers who learned LSE before 1 year of age from a family member; those participants who were not native LSE signers used sign language professionally, mostly as sign language interpreters. Proficiency in LSE was assessed with self-reported ratings on a scale from 1 to 5 (mean rating 4.64, SD 0.5). The second group was composed of 17 native speakers of Spanish with no knowledge of sign language. None of the participants had any knowledge of spoken Russian or Russian Sign Language. None had language, motor or neurological impairment and all reported normal hearing.

Participants were recruited in different parts of Spain, and all provided informed consent before the beginning of the experiment and were compensated for their participation. All participants underwent an MEG session, where they performed the experimental task, and a short MRI session to collect structural images (T1 and DTI – details below in section 4.2.4). Both sessions were performed at the Basque Center on Cognition, Brain and Language.

Out of 33 participants, three participants were excluded from the analysis due to technical errors during the MEG recording or very noisy data. One participant was excluded from the analysis because of very low performance in the behavioral task. The final pool of participants used for the analysis was of 14 controls (11 female, age: mean 31.5, SD 8.4) and 15 bimodal bilinguals (12 female; age: mean 39.4, SD 9.5).

4.2.2 Material

The material for the experiment consisted of videos of semi-spontaneous speech and sign in four different languages: Spanish, Russian, LSE and Russian Sign Language. For each language, two native speaker/signer models were recorded while retelling a short narrative based on a common set of comic strips with no or minimal language content. The overall set of stimuli comprised 320 videos, divided in 40 videos recorded by each of the eight models. The recording was performed at BCBL with a Kinect v2 (commercially known as Kinect for Xbox One,) custom-built set up which allowed the recording of both video and motion tracking information for each narrative. The motion tracking information is used to extract the frequency patterns from the visual signal in each language for the analysis. Models were recorded in front of a black background, in a light-controlled room. They familiarized themselves with the comic and then retold the story as if they were telling it to a friend for about one minute. Table 5 shows the mean and standard deviation of the video duration in seconds, across models and languages. A detailed explanation of the motion tracking set up and the stimuli creation can be found in section 3.3.

Table 5: Mean and standard deviation (in seconds) of the video-recording for each model and language.

	LSE # 1	**LSE # 2**	**RSL # 1**	**RSL # 2**	**Spanish # 1**	**Spanish # 2**	**Russian # 1**	**Russian # 2**
Mean (s)	63.83	59.68	59.22	56.54	61.12	62.63	62.74	58.97
Stand. Dev.(s)	4.06	3.84	4.56	2.60	2.21	2.57	2.85	2.10

To create material for the on-line task (a probe recognition task: see section 4.2.3 for a detailed explanation), we created two sets of clips: a set of probes extracted from the videos and a set of foils which were not part of the videos. For the probes we extracted a short five-second clip from each recorded video: the clip was manually selected to represent a perceptually salient moment of the video (for example, a striking gesture or a sign articulated at a particular location on the body). Each model also recorded ten extra vignettes in which they said or signed a list of specific expressions that would serve as foils: in spoken language we asked models to accompany their speech with emblems, that is, marked gestures with a specific meaning in their culture (Efrón, 1941); in sign language the models produced iconic signs accompanied by emotionally transparent facial features that are easily recognizable also by individuals with no knowledge of sign language. These recordings were then edited to create five-second clips.

All videos and probes were pre-processed with FFmpeg (version 2.7; Tomar, 2006): each video was cropped, the frame rate was adjusted to 25 fps, and the luminance was normalized across videos. A fade in and fade out of 4 frames (160 ms) was applied only to one-minute long videos. The audio track of spoken language videos was recorded with an external camera (Sony HDR-CX240E) at 48000 Hz sample frequency. The audios were normalized for sound level (70dB) using Praat (Boersma, P., & Weenink, 2020).

4.2.3 Procedure

During the MEG session participants viewed a total of 40 videos (ten videos for each language), divided in four consecutive language blocks. Each language block had five videos by each of the two models. The videos were distributed across the four blocks so that participants saw all 40 narratives once, and the narratives from each language model were shown the same number of times across all participants.

After the presentation of each video, participants saw two short clips of 5 seconds each (one probe and one foil). The clips were presented on the left or right side of the screen, and participants had to indicate which one of the clips was extracted from the video they had just seen. Responses were given with an MEG compatible response box, and the order and position of probe/foil presentation was counterbalanced across trials. This orthogonal task was designed to keep participants attentive throughout the experiment, especially during the unknown language blocks, and to serve as a filter to discard participants who were not looking at or paying attention to the screen throughout the session.

4.2.4 Data acquisition

MEG data were acquired at BCBL in a magnetically shielded room using the whole scalp MEG system (Elekta Neuromag, Helsinki, Finland). The system is equipped with 102 sensor triplets (each comprising a magnetometer and two orthogonal planar gradiometers), which are uniformly distributed around the head of the participant. Head position inside the helmet was continuously monitored using four head position indicator (HPI) coils. The location of each coil relative to the anatomical fiducials (nasion, left and right preauricular points) was defined with a 3D digitizer (FastrakPolhemus, Colchester, VA). This procedure is critical for head movement compensation during the data recording session. MEG recordings were acquired continuously with a bandpass filter at 0.01–330 Hz and a sampling rate of 1 kHz. Eye movements and cardiac rhythm were monitored with three pairs of electrodes in a bipolar montage placed on the external chanti of each eye (horizontal EOG), above and below right eye (vertical EOG) and on the left lower rib and below the left clavicle (ECG).

Continuous eye-tracking data were recorded during the MEG session with a ViewPixx TRACKPixx. Eye-tracking data were acquired at a 2000 Hz sample rate for both eyes. Additionally, all participants underwent an MRI single session, using a Siemens 3T MAGNETOM PRISMAfit at the BCBL. A T1-weighted (T1w) MRI and Diffusion Tensor Imaging (DTI) DTI scan were acquired. The structural brain data (T1w and DTI) and eye-tracking data are not discussed in the context of this doctoral thesis.

4.3 Analysis

4.3.1 Data preprocessing

Continuous MEG data were pre-processed off-line using the temporal Signal-Space-Separation (tSSS) method (Taulu & Simola, 2006) which suppresses external electromagnetic interference. MEG data were also corrected for head movements, and bad channel time courses were re-constructed in the framework of tSSS. Subsequent analyses were performed using MatlabR2012b (Mathworks, Natick, MA, USA) and Fieldtrip toolbox (Maris & Oostenveld, 2007). Heartbeat and EOG artifacts were detected using independent component analysis (ICA) and were linearly subtracted from recordings. The ICA decomposition was performed using the Infomax algorithm (Amari et al., 1996) implemented in Fieldtrip. Ocular and heartbeat ICA components were manually identified based on the spatial distribution and the temporal dynamics. Across participants, the number of heartbeat and ocular components that were removed varied from 0-1 and 0-3 components, respectively. Continuous MEG data were segmented into epochs of 4 seconds (with 2 seconds overlap), epochs with z-scores higher than 2.5 were considered as artifact-contaminated and rejected from further analysis.

4.3.2 Coherence analysis

We computed coherence to evaluate the phase synchronization between brain activity (MEG data) and the linguistic perceptual signals, speech for the spoken language and visual information for sign languages (Lizarazu et al., 2019; Molinaro et al., 2016; Molinaro & Lizarazu, 2018). In spoken language we extracted the speech envelope from the acoustic signal of each video. Envelopes of the stimuli were computed by applying the Hilbert transform to the auditory signals. For sign languages we selected the right hand speed vector as a proxy for the linguistic visual signal. The speed vectors were computed from the motion tracking data, and underwent the same preprocessing steps described in Chapter 3. Artifact free MEG data and the speech envelope were resampled to 25 Hz to match the sampling rate of the visual linguistic signals (Kinect data).

For each condition, coherence between the MEG segmented data and the relative linguistic signal (visual for sign languages and acoustic for the spoken languages) was calculated in the 0 – 12 Hz frequency band with 0.25 Hz frequency resolution. Coherence values of each gradiometer pair were summed.

The coherence bias was estimated empirically for each participant by randomly shuffling the original linguistic signals across segments, and re-calculating coherence in 100 permutations. For each sensor (combined gradiometers) coherence values were averaged together and then z-score transformed using the mean and standard deviation from the 100 permutations. Z-score transformations were calculated for each condition using the condition-specific mean and standard deviation from the random pairing dataset and with the same number of trials as the true pairing dataset.

Based on the previous literature on entrainment in spoken languages (Bourguignon et al., 2013; Destoky et al., 2019; Ghinst et al., 2019; Gross et al., 2013; Meyer & Gumbert, 2018; Molinaro et al., 2016; Molinaro & Lizarazu, 2018; Vander Ghinst et al., 2016) and a visual inspection of coherence plots, we selected two frequency bands of interest for our analysis. Delta band, from 0.5 to 2.5 Hz, is typically associated with prosodic rhythm in speech and it seems to largely overlap with relevant coherence in sign language based on a previous study (Brookshire et al., 2017). Theta frequency band, from 4 to 7 Hz, is linked with syllabic rate and it has often been identified in spoken language cortical tracking. The mean of the z-scored coherence values was obtained in each frequency band (delta and theta) across all channels for each participant and condition.

The analysis was organized in the following manner. First, to examine the effects of language knowledge and modality, we ran a 2 (language modality) x 2 (language familiarity) analysis for the bimodal bilinguals. The aim here is to look at the effect of each factor (and their possible interaction) on entrainment, separately for each frequency band. We expect to find differences in entrainment between known and unknown languages in both modalities (Spanish vs Russian and LSE vs RSL). When comparing entrainment across modalities we expect to find two differences. On the one hand, we anticipate a difference in the topography of coherence, as a function of the sensory cortex recruited. On the other hand, we expect a difference in the frequency bands involved, based on the temporal structure of the linguistic input: while speech involves both delta and theta entrainment, we expect entrainment to sign language to be limited to delta band. Second, we examined the same contrasts in the control group. When comparing Spanish and Russian we expect to find the same results as in bimodal bilinguals participants; conversely when comparing LSE and RSL we do not expect to find any differences in entrainment since both signed languages are unknown to these participants. In terms of modality, the contrast between Spanish and LSE is not useful since it confounds both modality and language knowledge in this group, but we can compare Russian and RSL to isolate the effect of modality on entrainment. Again, we expect to find similar results to those of the bimodal bilinguals. Finally, we compared the findings for each group to see whether bimodal bilinguals and control participants had similar or different patterns of entrainment. In line with our predictions, we expect the groups to have similar patterns for the spoken languages but different patterns for the signed languages. Additionally, we directly compared bimodal bilinguals versus controls for each modality, looking at each frequency band separately. The comparison of the entrainment to either speech or sign in bimodal bilinguals and controls reveals whether knowing a sign language has an impact on how each type of language signal is perceived.

For each comparison, we performed cluster-based permutation tests (Maris & Oostenveld, 2007) to assess statistical differences in coherence values in specific frequency bands (delta: 0.5 – 2.5 Hz, theta: 4 -7 Hz). This non-parametric permutation analysis allows us to avoid the problem of multiple comparisons among the high number of sensors. Briefly, clusters of channels with significant differences ($p < 0.025$) were created by spatial adjacency (at least two neighbouring channels). The neighbourhood definition was based on the distribution of the MEG sensors (combined gradiometers). A set of 1000 permutations was created by randomly assigning condition labels and then t-values were computed for each permutation. A cluster was considered to have a statistically significant effect if the sum of t-values in the original dataset was greater than the 95th percentile ($p < 0.05$) of the distribution of the corresponding values in the randomized data.

To examine the interaction of two effects, the difference between coherence in two conditions of one effect (e.g., known and unknown language for language knowledge) was first calculated separately in each of the conditions of the other effect (e.g., speech and sign for modality), and then these two differences were compared with a cluster-based permutation analysis. If this comparison yields any difference it means that there is an interaction between the two effects, and thus motivates post hoc contrasts in subsequent analysis. In the absence of an interaction effect instead we can collapse the two conditions in each variable in turn to investigate main effects.

4.4 Results

4.4.1 Behavioral data

During the experiment participants performed an orthogonal task with a double aim: to keep them attentive throughout the task and to filter participants who were not paying attention to the videos presented on the screen. Figure 16 shows accuracy mean for each participant across all four languages (Spanish, Russian, LSE and RSL). All participants but one show high accuracy in all conditions, with results almost at ceiling for both known and unknown languages. One participant belonging to the control group performed below chance in all four languages; this result suggests that this participant was not paying attention to the videos presented during the task, and was therefore excluded from subsequent analysis.

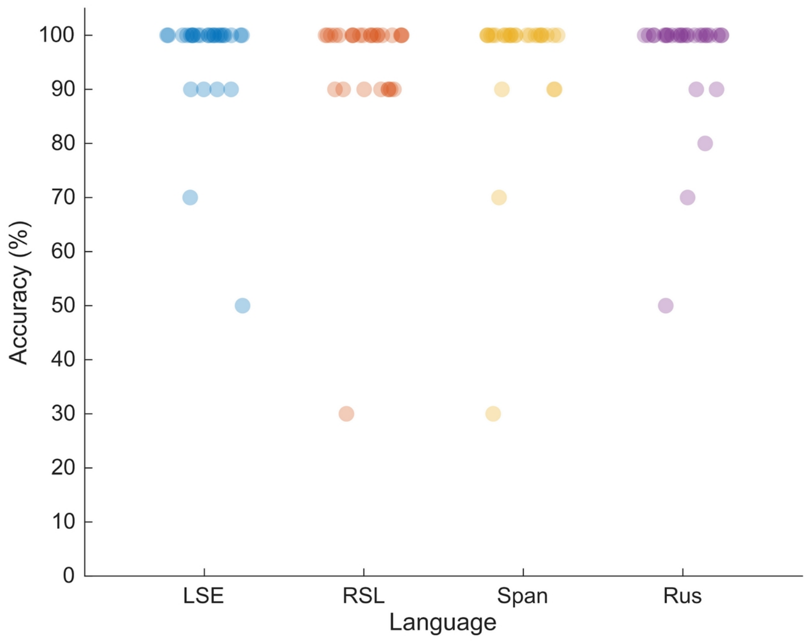

Figure 16: The scatter plot shows the accuracy on the orthogonal task performed in the MEG. Dots represent the accuracy for each participant in each language.

4.4.2 MEG data

4.4.2.1 Bimodal bilinguals

We first tested the possible interaction between language knowledge and language modality in bimodal bilinguals. This analysis revealed an interaction effect (p = 0.002) between language knowledge and modality in the delta frequency band; while no interaction is found in the theta (Figure 17). Based on these results we can investigate simple effects of language knowledge and modality in the delta frequency band. In the theta band, due to the lack of an interaction, we only investigate main effects.

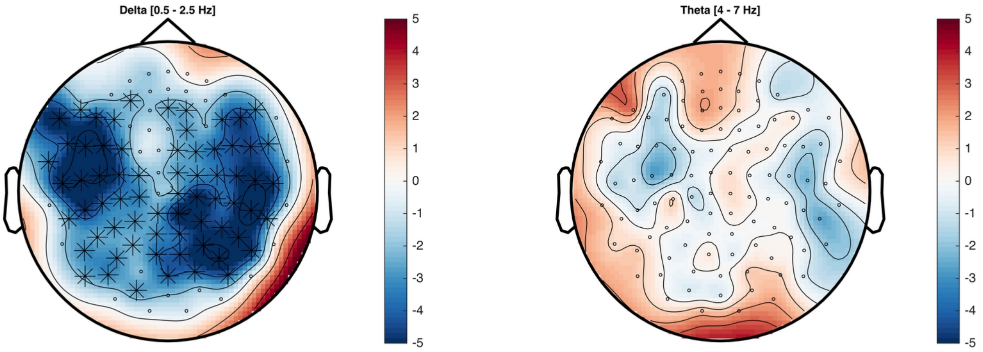

Figure 17: Plots showing the interaction analysis for bimodal bilinguals participants (n = 15). First the difference in coherence between known and unknown language was calculated in each modality (Spanish vs Russian and LSE vs RSL). The plots show the difference of these values between spoken and sign modality in delta (left), but not in theta (right) frequency bands.

Language knowledge

Delta: When comparing Spanish and Russian in the delta frequency band we see a widespread cluster of sensors (p = 0.002) showing more coherence for Russian than Spanish. The cluster is located bilaterally over temporal areas, in line with the topography associated with entrainment to spoken languages (Figure 18, left). The comparison between LSE and RSL shows more coherence for LSE than RSL in a cluster of sensors located in the right hemisphere (p = 0.002), as shown in Figure 18 (right).

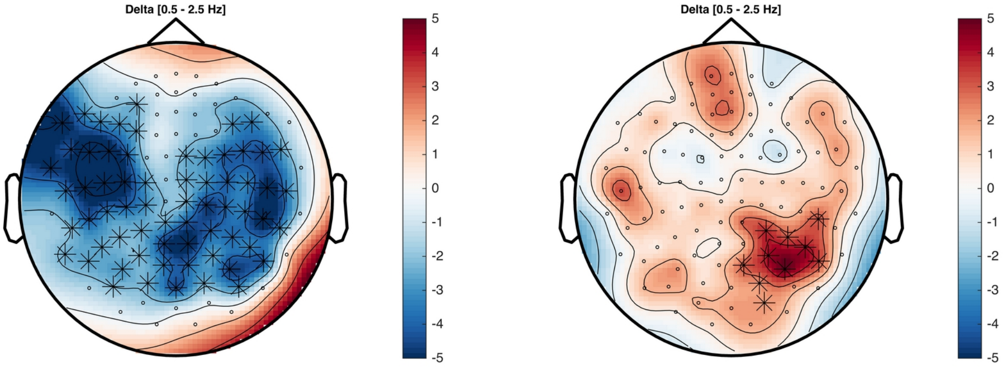

Figure 18: Plots showing the difference in coherence between Spanish and Russian (left) and between LSE and RLS (right) in delta frequency band in bimodal bilinguals (n = 15).

Theta: The comparison between known (Spanish and LSE) and unknown (Russian and RSL) languages (Figure 19) revealed a cluster of significant sensors located in right parietal areas, showing more coherence for unknown languages compared to known languages.

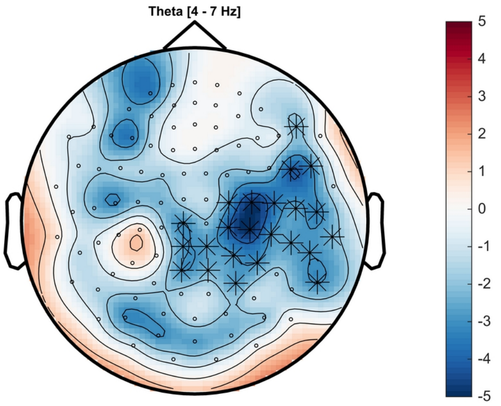

Figure 19: Plots showing the difference in coherence between known and unknown languages in theta frequency band in bimodal bilinguals (n = 15).

Language modality

Delta: When comparing Spanish and LSE in the delta frequency band we see a bilateral temporal cluster showing more coherence in spoken language compared to sign language (p = 0.008). Additionally, a marginally significant cluster (p = 0.07) shows more coherence for sign than spoken language, and this cluster is located over right parietal regions (Figure 20, left). The comparison between Russian and RSL highlights a bilateral temporal cluster showing more coherence in spoken language compared to sign language (p = 0.002). No regions showed significantly more coherence for sign than spoken language coherence (Figure 20, right).

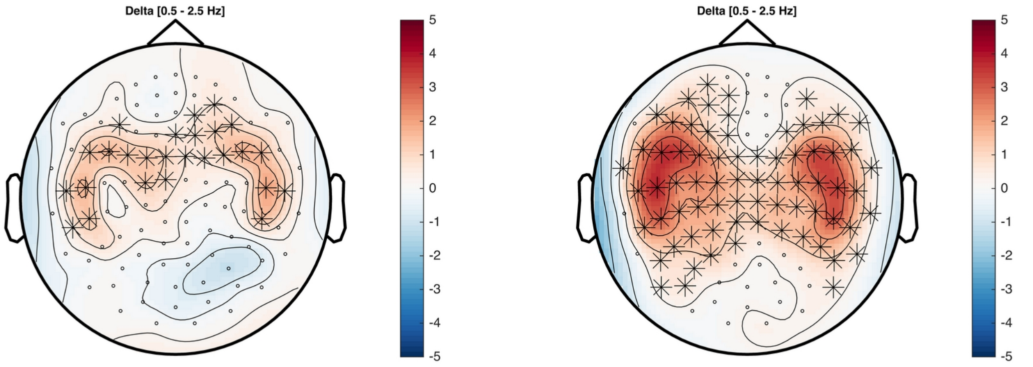

Figure 20: Plots showing the difference in coherence between Spanish and LSE (left) and between Russian and RLS (right) in delta frequency band in bimodal bilinguals (n = 15).

Theta: The comparison of spoken (Spanish and Russian) and signed (LSE and RSL) languages (Figure 21) shows more coherence in spoken language compared to sign language, in a widespread bilateral cluster of sensors (p = 0.002).

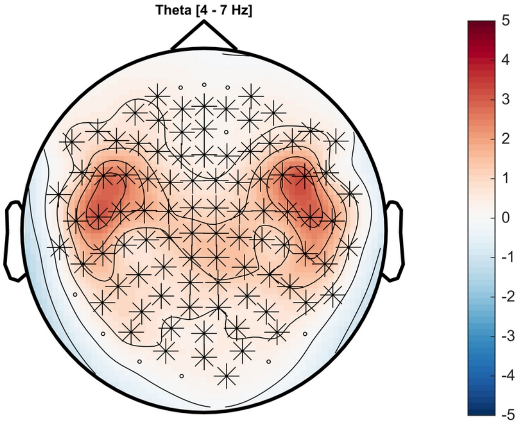

Figure 21: Plots showing the difference in coherence between spoken and signed languages in theta frequency band in bimodal bilinguals (n = 15).

4.4.2.2 Controls

The analysis of the interaction between the two main effects revealed an interaction effect (p = 0.002) in the delta domain; no interaction is found in the theta domain (Figure 22). Therefore, we investigate simple effects between the different languages in the delta frequency band. In the theta band, the lack of an interaction limits the analysis to main effects, but these are not meaningful and do not offer a straightforward interpretation. The lack of knowledge of LSE in this group gives rise to an imbalance of known and unknown languages in the comparisons (Spanish+LSE versus Russian+RSL; Spanish+Russian versus LSE+RSL). For the sake of completeness, these main effect contrasts are reported in Appendix 3.

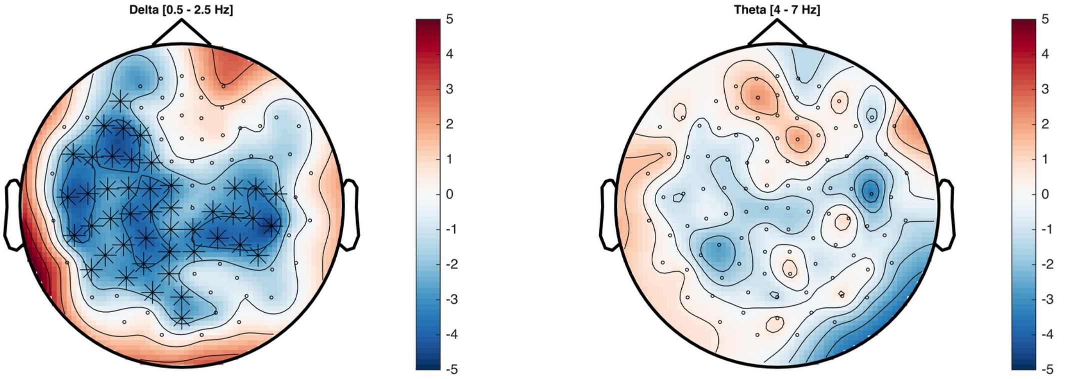

Figure 22: Plots showing the interaction analysis for control participants (n = 14). First the difference in coherence between known and unknown language was calculated in each modality (Spanish vs Russian and LSE vs RSL). The plots show the difference of these values between spoken and sign modality in delta (left), but not in theta (right) frequency bands.

Language knowledge

In order to investigate the effect of different languages on control participants we separately compared spoken languages (Spanish vs Russian) and signed languages (LSE vs RSL) in delta frequency band.

Delta: The comparison between known and unknown spoken languages (Figure 23, left) shows a bilateral cluster indicating more coherence in Russian compared to Spanish (p = 0.002), similar to what was found in bimodal bilinguals. When comparing LSE and RSL, we do not find any significant difference in coherence (Figure 23, right).

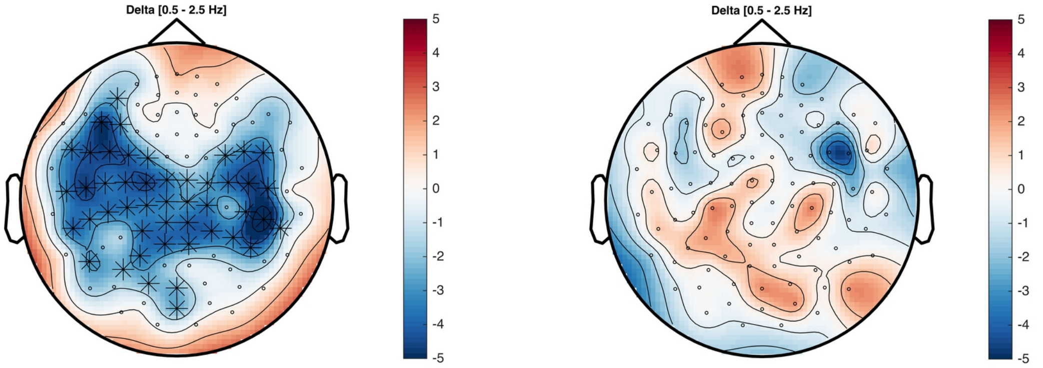

Figure 23: Plots showing the difference in coherence between Spanish and Russian (left) and between LSE and RLS (right) in delta frequency band in controls (n = 14).

Language modality

Since the control participants did not know LSE, the contrast between Spanish and LSE confounds both modality and language knowledge for this group (this contrast is reported in Appendix 3, Figure A4). Therefore, when comparing across language modality we are focusing only on Russian and RSL (both unknown languages).

Delta: The comparison between Russian and RSL (Figure 24) shows a bilateral temporal cluster with more coherence in spoken language compared to sign language (p = 0.002). No regions showed significantly more coherence for sign than spoken language.

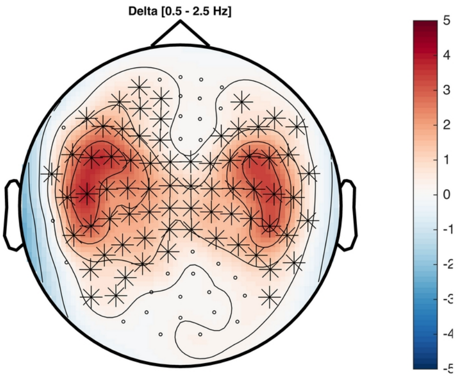

Figure 24: Plots showing the difference in coherence between Russian and RSL in delta frequency band in controls (n = 14).

4.4.2.3 Comparison between participant groups

To assess possible differences between bimodal bilinguals and controls in the patterns of entrainment based on language knowledge, for each modality we examined the interaction between participant group and language knowledge. These analysis was limited to the delta band, since previous comparisons showed that theta frequency band is not relevant. For speech (Figure 25, left) there was no interaction, indicating that there was no significant difference in the pattern of entrainment to Spanish compared to Russian between the two groups. In contrast, for sign (Figure 25, right), there was an interaction between group and modality (p = 0.008), revealing that the pattern of entrainment to LSE compared to RSL was significantly different between the two groups.

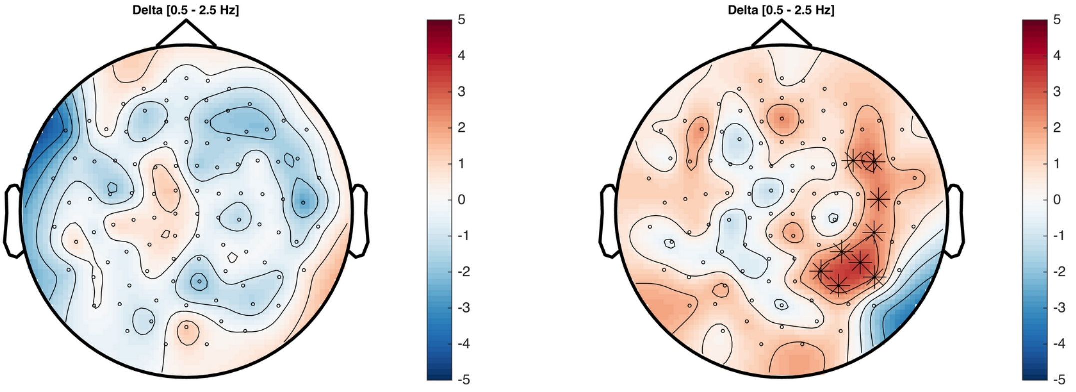

Figure 25: Plots showing the interaction analysis between language knowledge and participant group in the delta band for spoken languages (left) and signed languages (right).First the difference in coherence between known and unknown language was calculated. The plot shows the difference of these values between bimodal bilinguals and controls.

For the direct comparisons between participant groups, we compared the coherence between bimodal bilinguals and controls for each modality separately by collapsing coherence data from the two respective languages (Russian and Spanish; LSE and RSL).

Sign modality

Figure 26 presents the comparison between bimodal bilinguals and controls in sign languages.

Delta: Bimodal bilinguals showed higher coherence compared to controls (p = 0.004) in a cluster of sensors located in the right parietal cortex.

Theta: There was no significant difference in coherence between participant groups.

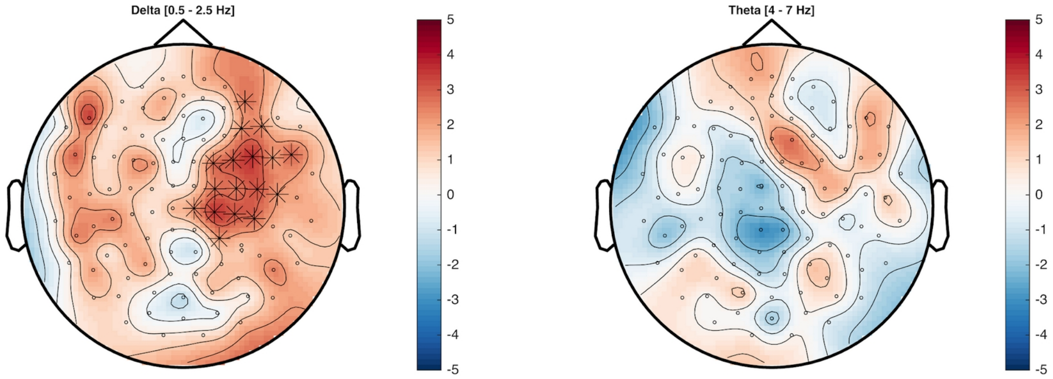

Figure 26: Plots showing the difference in coherence in the sign modality between bimodal bilinguals and sign-naive participants in delta (left) and theta (right) frequency bands.

Spoken modality

Figure 27 presents the comparison between bimodal bilinguals and controls in spoken languages.

Delta: There was no significant difference in coherence between participant groups.

Theta: Bimodal bilinguals showed stronger coherence than controls in a distributed cluster of centro-parietal sensors (p = 0.03). These results suggest that knowing a sign language affects not only entrainment to sign language, but also to spoken language.

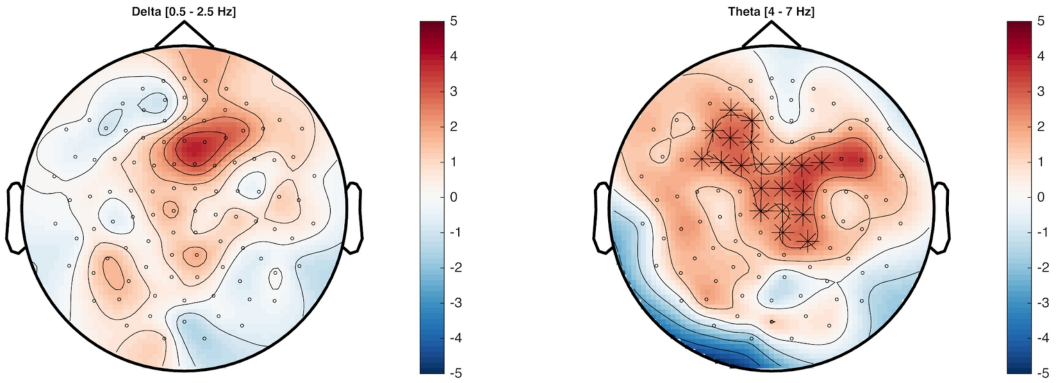

Figure 27: Plots showing the difference in coherence in the spoken modality between bimodal bilinguals and sign-naive participants in delta (left) and theta (right) frequency bands.

4.5 Discussion

4.5.1 Sign language

The main goal of this study was to investigate whether language-brain entrainment is limited to spoken language, or is recruited to process sign language as well. One previous study found entrainment between the visual signal extracted from sign language videos and brain activity recorded with EEG (Brookshire et al., 2017). Results from our experiment confirm these findings. We see that bimodal bilinguals show more coherence than sign-naive participants when presented with videos in sign language (Figure 26). Interestingly, this effect is only present in the delta frequency band (0.5-2.5 Hz), while no difference is found in theta (4-7 Hz). The relevance of the delta band to sign language is confirmed by the comparison between LSE and RSL, the known and unknown (sign) language for the bimodal bilinguals: these participants entrained more to LSE in delta band only (Figure 18); conversely controls did not show any difference (Figure 23). This results was confirmed by the difference between the two groups shown in Figure 25.

Taken together, these results show that sign language elicits language-brain entrainment, but the characteristics of entrainment are modulated by some specific properties of the sign modality. If we analyse the frequency make up of the sign language stimuli used in this experiment, we can see in Figure 28 that the periodic components of both LSE and RSL are concentrated in low level frequencies spanning between 0.5 and 2.5 Hz. This temporal grain fits well with the overall larger time scale associated with sign language, which is constrained by the size of the sign language articulators. Sign rate production falls around 2 Hz and sign duration is about twice that of a monosyllabic word (Grosjean, 1977; Klima, E. S., & Bellugi, 1979; Wilbur, 2009). Higher frequency bands, such as theta, do not seem to play a relevant role in sign production and processing. As a caveat, it is important to notice that rhythmicity associated with higher frequencies might emerge when looking at smaller sign language articulators such as the mouth, eyebrows and fingers. The temporal resolution of processing is not the only difference between entrainment in spoken and signed languages. Coherence in LSE occurs in sensors located over the right angular gyrus. This brain area has been associated with biological motion processing (Allison et al., 2000; Puce & Perrett, 2003), and linked with sign language processing (Emmorey, 2021; Levänen et al., 2001). The comparison of Spanish and LSE in bimodal bilinguals (Figure 20) reveals a clear topographical dissociation between entrainment to spoken and sign languages: spoken language processing recruits auditory cortices in both hemispheres, not overlapping with areas activated by sign language. The topography of sign language entrainment in this study does not overlap with occipital and frontal regions found by Brookshire (2017), but it does fit well with literature on language brain networks for sign language processing. Spoken and sign language seem to rely largely on the same network of brain areas located around the perisylvian cortex (Emmorey, 2021), but show the activation of modality-specific areas during language processing. In sign language the superior parietal cortex is linked with motion analysis during language comprehension.

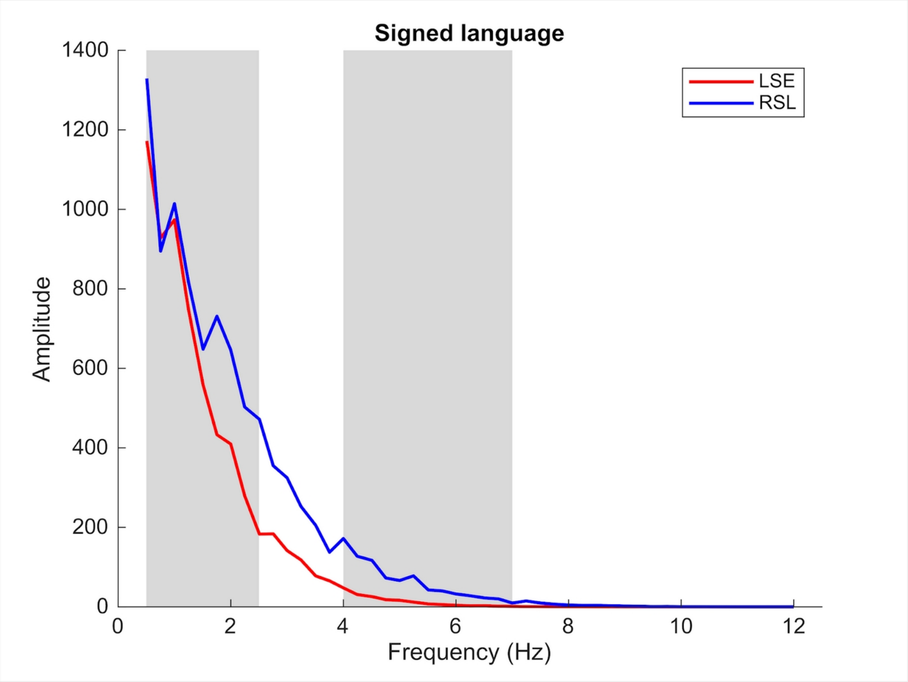

Figure 28: Power spectrum of the right hand speed vector for LSE (red) and RSL (blue). Shaded areas highlight the two frequency bands of interest: delta (0.5 - 2.5 Hz) and theta (4 - 7 Hz).

4.5.2 Spoken language

Our study reproduces the classical findings of speech-brain entrainment found in the literature. Both Spanish and Russian show coherence in delta frequency band (0.5-2.5 Hz) and theta frequency band (4-7 Hz). These frequency bands have been extensively linked with different linguistic features: theta is associated with syllable parsing (Ding et al., 2017) while slower delta oscillations coincide with prosody patterns (Bourguignon et al., 2013; Keitel et al., 2017). Coherence in spoken languages is located in bilateral temporal regions overlapping with auditory processing areas.

The comparison of entrainment between known (Spanish) and unknown (Russian) languages revealed unexpected results: participants entrained more to Russian compared to Spanish, especially in the delta frequency band. This result holds true for both bimodal bilinguals (Figures 18, 19) and controls (Figure 23). We predicted more entrainment to Spanish, as our participants are native speakers of this language and fully understand it, while Russian is completely unknown to them. One possible explanation for this result is the difficulty of the task: in order to properly perform the orthogonal task for the trials in Russian participants might have allocated extra attention to these videos, and the enhanced entrainment could be a consequence of a top-down attentional effect. Indeed, the behavioral results show that participants' performance is almost at ceiling in both spoken languages, without any clear difference between Russian and Spanish. Moreover, during audio-visual speech presentation of a known language attention usually focuses on the eye region, while with unfamiliar or unknown languages the gaze patterns shift towards the mouth (Barenholtz et al., 2016). It is possible that participants deployed more attention to the mouth region during Russian video presentation and therefore entrained not only to the acoustic component of spoken language, but also to the visual signal generated by mouth movements. Previous studies found that mouth movements during speech production oscillate at 4-5 Hz, and these frequencies overlap with speech envelope periodicity (Walsh & Smith, 2002). Overt attention to lip movements during speech increases coherence between frequency of lip movements and the recorded MEG brain signal (Park et al., 2016). Another possible explanation of the difference we find in coherence for Spanish and Russian lies in the temporal structure of these two languages. Spanish is a syllable-timed language, in contrast with Russian, which is characterized by a stress-timed organization (Abercrombie, 2019; Nespor et al., 2011). Both types of language exhibit periodicity focused in the classical delta and theta frequency bands, but stress-timed languages, such as Russian, are known to have less power in theta and more power in delta since the rhythm patterns follow the stressed syllables. Again, we can examine the average power spectrum extracted from the speech envelope of Spanish and Russian videos used in the experiment (shown in Figure 29): Russian is characterized by increased power in delta band in line with its stress-timed structure. Higher coherence in Russian than Spanish therefore might be partly driven by this difference in power.

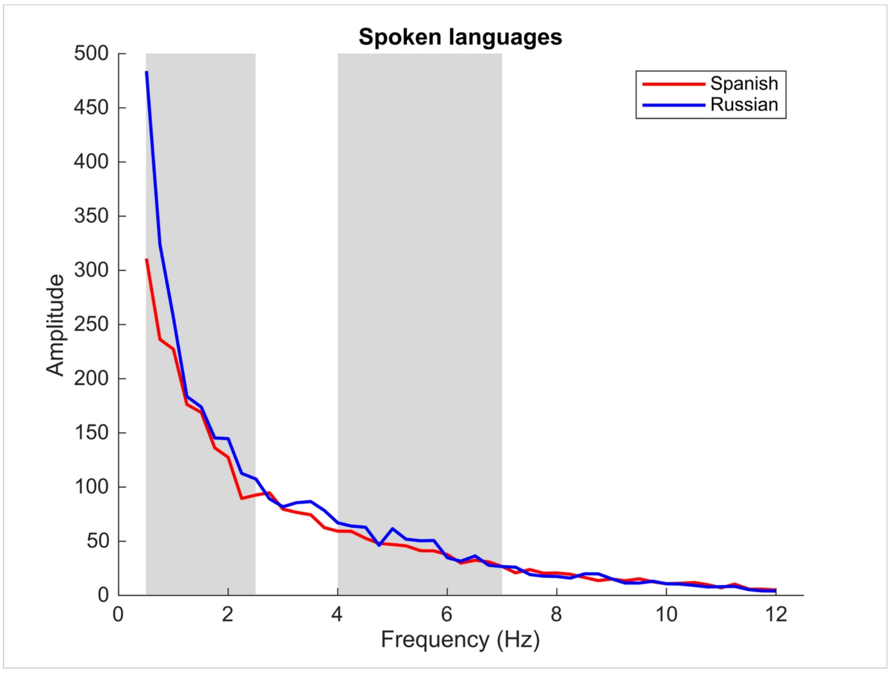

Figure 29: Power spectrum of the speech envelope for Spanish (red) and Russian (blue). Shaded areas highlight the two frequency bands of interest: delta (0.5 - 2.5 Hz) and theta (4 - 7 Hz).

The experiment design allowed us to investigate the effect of experience with sign language on spoken language processing. We found that bimodal bilinguals entrain more than controls to spoken languages, and this difference is restricted to the theta frequency bands. Knowledge of a sign language might aid in recruiting the visual information in the spoken language videos, such as gestures and mouth movements. If bimodal bilinguals make a greater use of the visual information that accompanies speech compared to controls, we should be able to see this difference when analyzing entrainment to the visual component of the spoken language signal. This result is particularly interesting as it provides evidence for a cross-modal transfer effect between sign and spoken language: knowing a sign language may change how you perceive and comprehend spoken language.

4.5.3 Comparing spoken and sign language

The comparison between results in spoken and signed languages help us identify those properties of language-brain entrainment that are independent of the temporal structure and the modality of language. Overall, our results show that both spoken and signed languages rely on language-brain entrainment during language processing, although with some differences. Frequency bands and topography of entrainment differ between signed and spoken languages (see section 4.5.1 and section 4.5.2, respectively), in line with the specific properties of each language. Interestingly, spoken language processing elicits language-brain entrainment to a greater extent than sign language processing. This is clear when comparing spoken and sign language: coherence is overall much higher for acoustic compared to visual input (see Figures 20, 21 for bimodal bilinguals; Figure 24 for controls).

This effect might be due to the inherent difference in periodicity in the linguistic signal: spoken languages compared to sign languages are characterised by higher power in coherence in all frequencies, suggesting that they have more pronounced periodic components (Figure 30). An important caveat is that we are comparing speech, which represents the whole linguistic acoustic input of spoken language, with the signal from one single articulator in sign language. The right hand speed vector does not represent the full temporal properties of the visual signal, and an aggregated measure of various articulators could prove more informative. Another possibility is that the auditory perceptual system is more sensitive to temporal regularities in the incoming signal, while the visual system favours spatial over temporal processing. The natural predisposition of the auditory system to use temporal patterns in the signal could drive the increased entrainment. Finally, this imbalance in the strength of entrainment between the two modalities could arise from the design of our study. For spoken languages we presented audio-visual stimuli and therefore participants were exposed to both the auditory (in form of speech) and visual (in form of gestures) signals at the same time, while in the sign modality the visual signal was the only linguistic information presented. Co-speech gestures are known to couple with the rhythmicity of the acoustic speech signal (Wagner et al., 2014), resulting in higher power of periodicity. In the same vein, presentation of congruent audio-visual speech enhances entrainment compared to auditory only speech (Crosse et al., 2015).

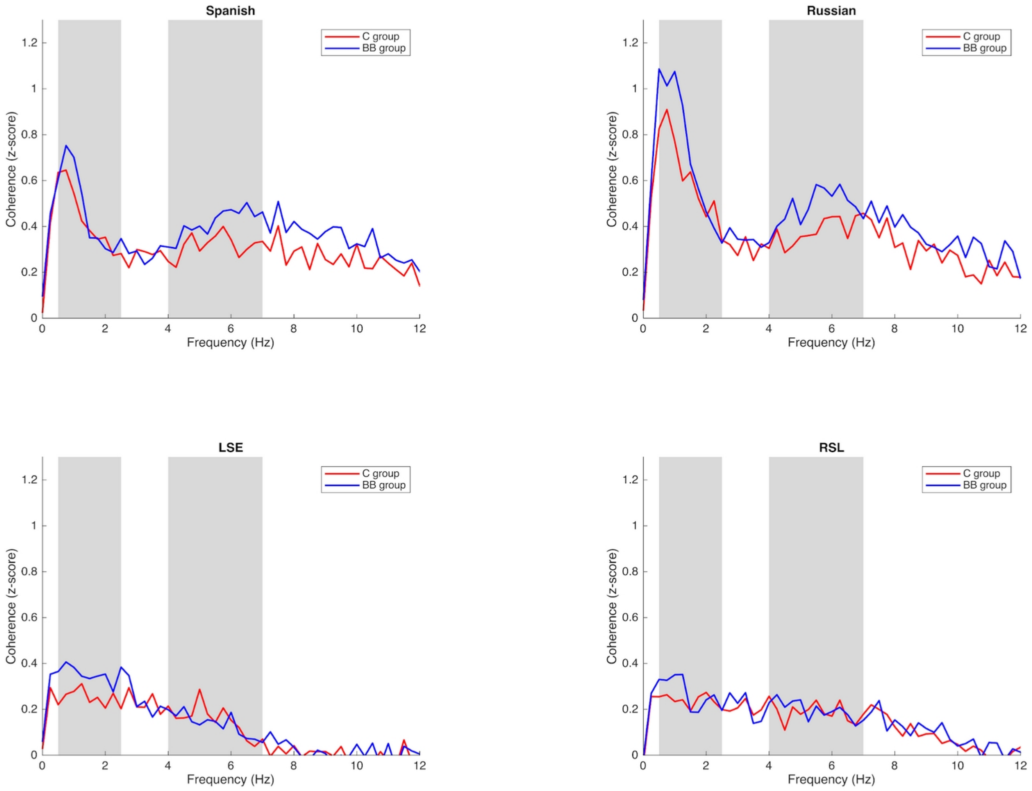

Figure 30: Coherence between MEG recording and linguistic signal: speech envelope for spoken languages and right hand speed vector for signed languages. The plots present z-score coherence values for Spanish (top-left), Russian (top-right), LSE (bottom-left) and RSL (bottom-right). In each language values are plotted separately for participants in the control group (C, in red) and bimodal bilinguals (BB, in blue). Shaded areas highlight the two frequency bands of interest used for subsequent analysis: delta (0.5 - 2.5 Hz) and theta (4 - 7 Hz).

4.6 Conclusion

This study represents one of the first attempts to reproduce language-brain entrainment with sign language, and therefore it allows us to characterize this phenomenon by abstracting it from the properties of the specific modality used to perceive the signal. The analysis presented in this chapter focuses on two specific components of the multidimensional signal of spoken and sign languages: the acoustic speech signal and the visual right hand signal. Previous studies showed that co-speech gestures and mouth movements do play a role in entrainment (Crosse et al., 2015; Park et al., 2018), and our results hint that the use of the visual information in spoken language might be modulated by experience with a sign language. Using information theory analysis could help us to better disentangle the different shared and individual contribution of auditory and visual signals

(Park et al., 2018). Just as speech is a multidimensional (and multimodal) signal, in sign language the linguistic information is conveyed by multiple visible articulators; the same information theory approach can be applied in this context to investigate which articulators are relevant to sign language processing and their specific contribution. Analysis of the full temporal spectrum of sign language will help understand how this multidimensional signal is combined for processing of the whole linguistic input.

The results also provide evidence that language knowledge impacts entrainment. We find differences in entrainment between known and unknown languages, but the direction of the effect depends on the modality: in the spoken domain the unknown language elicits more entrainment than the known language, while in sign language the pattern is reversed. These findings do not provide a definite answer to the debate surrounding the effect of language knowledge on entrainment. The different temporal structure of two spoken languages compared in this study – syllable-timed for Spanish and stress-timed for Russian – may account for some of the differences that we find. The higher coherence we see in the unknown spoken language might be driven by its rhythmic pattern, or by an extra attentional load due to the novelty or difficulty of the language. Future work that controls for the timing pattern of the unknown language and modulates task difficulty could rule out these possible explanations and identify the specific contribution of language knowledge to entrainment.

Overall, the results of this study contribute to a better characterization of the role of entrainment in language processing. We find evidence that entrainment is a modality-independent mechanism: phase synchronization between the oscillations of neuronal populations and the temporal regularities in the physical signal is a common process to decode both acoustic and visual linguistic information. Rhythmic patterns in language are exploited by our cognitive system to anchor an optimal processing, and ultimately comprehension, of language. Nevertheless, entrainment depends on the rhythmic patterns of the input and this, in turn, is modulated by intrinsic properties of the perceptual modality. Sign language shows entrainment in the delta frequency band, associated with slower periodicity of the right hand movement, and its topography is specifically located in a region devoted to motion processing. Another important difference between spoken and sign languages is the strength of entrainment, with higher coherence in the spoken compared to the sign domain. This result suggests that sign language processing does not rely on the temporal periodicity of the linguistic signal as much as spoken language does; this idea is further developed in section 5.2.2.

Chapter 5: General discussion

In this chapter I revisit the research questions laid out in section 1.5 and answer them in light of the main findings of the studies presented in this doctoral work. I first summarise the results from each study and then I lay out the theoretical implications of these findings and link them to the general framework of language processing. In section 5.2.1 I focus on the relationship between modality and the temporal structure of the signal during language production, and specifically answer to the questions:

- Can we distinguish the temporal patterns of the visual signal in sign language and spoken language? If so, which kinematic properties are better suited to describe the signal?
- Do different sign language articulators show distinct temporal patterns?

The way our cognitive system perceives and processes the temporal structure of the signal is covered in section 5.2.2. This section addresses the following questions:

- Does the temporal structure of language play a role in language comprehension?
- Does modality affect this relationship between temporal structure and language comprehension?
- Is language-brain entrainment recruited for sign language processing?
- Does the specific temporal structure of sign language modulate the characteristics of entrainment?

5.1 Summary of findings

This PhD thesis investigated how the modality used to produce and perceive a language affects the temporal structure of language itself and how it is processed by our cognitive system. Sign language represents the perfect test case to disentangle the effect of modality: sign languages share the same linguistic complexity as any spoken language but employ solely the visual modality. In this work we performed three studies to compare the temporal structure and processing of Spanish and Spanish Sign Language (LSE) from different points of view.

The experiment presented in Chapter 2 focused on how language intelligibility is disrupted by distorting the temporal structure of language. We used a locally reverse-time speech paradigm, adapted to both the visual and acoustic modality, to apply different levels of distortion to Spanish, LSE and visual temporally-structured, non-linguistic visual stimuli. Our results showed a marked difference in the pattern of intelligibility loss between spoken and sign language: in Spanish intelligibility was completely lost after a certain level of distortion; in LSE, instead, the decrease in intelligibility was slow and constant without any clear perceptual bottleneck. Moreover, in LSE, even with the highest level of distortion, intelligibility of the signs was still quite high (and plateaued at a level at which more than half of the signs were correctly perceived). The non-linguistic visual signal showed a similar gradual and constant pattern, but temporal manipulation had a greater impact on this signal compared to LSE. Modality poses some constraints on the way distorted information can still be retrieved and processed, and the presence of linguistic structure aids the decoding of information.

In Chapter 3 we investigated the kinematic properties of the visual signal in both spoken and signed languages, and explored different measures that can describe and distinguish these two types of signal. We used a Kinect motion tracking system to record 3D coordinates of different body and face points of native signers and speakers during naturalistic storytelling. Several kinematic measures (such as motion magnitude and space, number of submovements, and UMAP clustering based on time-frequency profiles) showed reliable differences between language modalities. These exploratory analyses demonstrate that sign language, compared to the visual signal that accompanies speech, is characterized by more periodic and homogenous kinematic patterns and by a bigger use of the space. Interestingly, our results showed that not all sign language articulators show similar characteristics, pointing to the importance of disentangling the specific contribution of each articulator.

Chapter 4 presented an MEG study investigating the phenomenon of language-brain entrainment in sign language. We modulated language modality and language familiarity by presenting videos of natural storytelling in Spanish, Russian, LSE and RSL to bimodal bilinguals and sign-naive participants while their brain activity was recorded with MEG. We found that the bimodal bilinguals' brain activity synchronized with both the speech envelope in spoken language and the movements of the right hand in sign language. The characteristics of entrainment vary depending on the modality: in spoken language we reproduced the classical results found in the literature, namely, entrainment in delta and theta frequency bands over bilateral temporal regions; in contrast, sign language entrainment is restricted to the slower delta frequency band, in line with the periodicity of the visual signal, and located over right parietal areas associated with motion processing. Of note, entrainment in sign language was much weaker compared to what is found in spoken language. Language knowledge impacted entrainment, but the nature of the relationship is unclear: in sign language LSE elicited more entrainment than RSL, but the pattern is reversed in the spoken domain. Finally, we find evidence that suggests that experience with a sign language influences the way we entrain to the spoken audio-visual signal.

5.2 Theoretical implications

This thesis examines the importance that temporal structure has in language production and processing. Time represents the filter through which we perceive any internal or external event, and this perception depends on two distinct but interconnected elements: the temporal structure of the event itself and the perceptual system used by the brain to perceive it. These two elements can be thought of as pieces of a puzzle, and full and effortless perception can be achieved only when the two pieces fit together. The form of these pieces is not fixed, but is shaped by several factors. Here we examine one of the most important factors, namely the signal modality.

The acoustic and the visual systems have been extensively investigated as they represent the main gateways to perception. The time dimension is understudied in vision, compared to the attention it has received in auditory processing. This is mainly due to the importance temporal structure plays in the acoustic domain, while vision relies more on spatial information. In the specific case of language, though, time structure seems to play a relevant role for language processing, calling for a better characterization of temporal patterns in sign language.

5.2.1 Temporal structure of the signal in language production

The modality used to produce the linguistic signal, visual in the case of sign language and acoustic in the case of spoken language, deeply affects the nature of the signal itself. The visual domain favours the presentation of simultaneous but spatially distinct sources of information. In sign language this is evident in the use of several distinct articulators, such the hands, the torso, the head and facial articulators. This makes the sign language a multilayered signal: we can isolate the information coming from each articulator (or layer). Speech can also be thought of as a complex signal containing multiple levels of information (e.g. pitch, prosody, rhythm) but we perceive it as one temporally evolving signal, the result of an 'invisible' aggregation process of the combined motion of different vocal tract articulators, many of which are not visible to the receptor. This represents a fundamental difference between the speech acoustic signal and visual (multilayered) sign language signal.

In Chapter 3 we saw that motion tracking analysis can capture the temporal dynamics of the visual signal in sign language and, importantly, that this dynamic is different across different articulators (the analysis focused on five: head, torso, right hand, right shoulder and left hand). The temporal pattern of the right hand and shoulder, for example, reliably distinguish between whether the person is producing sign language or spoken language, suggesting that these two articulators play a pivotal role in conveying linguistic information. In contrast, other articulators, such as head and torso, move similarly in both spoken and sign languages. This finding supports the importance of multilayered nature of sign language: each articulator not only carries different linguistic content (Sandler, 2018), but it also displays different kinematic properties compared to the movement of other body parts. Each articulator represents one single layer of the complex visual signal created during sign language production, but the linguistic input is perceived as a whole. Similarly to what happens in spoken language, the visual signals coming from each articulator combine together to create one physical signal processed by our cognitive system. The nature of this combination is still unclear. We know that different articulators show different temporal patterns, but more research is needed to understand what the relative contribution of each articulator is to the perception of visual language.

When comparing the visual signal of sign language and the visual signal that accompanies speech we find that, even if produced with the same articulators and in the same modality, they show different kinematic features. This result proves that the visual modality itself is not the sole driving force behind the characteristics of the signal, but interacts with other properties of the signal. Spoken language can be multimodal when both speech and hands, body and face movements are produced simultaneously to convey information. This visual component has been shown to contribute semantic, syntactic, discursive and pragmatic information to the verbal (i.e., auditory) part of an utterance (Arnheim & McNeill, 1994; Kendon, 2015; McNeill, 2013), but it is still subordinate to speech. In spoken language, speech is generally sufficient to convey all the information, and body and face movements aid communication but are not necessary for it (otherwise the telephone and the radio would never have become as ubitquitous as they have). The dominant role played by the auditory channel in speech is also clear when looking at the physical properties of the acoustic and visual signal in spoken language: the temporal pattern of body movements tends to couple with that of speech (Danner et al., 2018; Pouw & Dixon, 2019; Wagner et al., 2014b). The bodily movements that accompany speech accommodate the temporal properties of the acoustic modality employed by speech. In contrast, sign language employs solely the visual modality and the signal organizes naturally according to the affordances of the visual system.

The results presented in Chapter 3 demonstrate the validity of motion tracking data and kinematic analysis in the study of sign language to better understand the organization of the linguistic information. Our findings also advocate for the use of MOCAP technologies which spatially isolate different parts of the signal. Systems like Kinect and OpenPose have the spatial resolution to separately record and measure the movements of each single articulator. Pixel differentiation methods, such as IVC and Optical Flow, are certainly useful to give an aggregate measure of sign language visual signal but more research is needed to to understand whether collapsing all information layers into one stream accurately reflects the way we process the sign language signal at the cognitive level.

5.2.2 Temporal structure of the of the signal in language processing

In order to be understood, the linguistic signal needs to be processed by our cognitive system. Our brain has different perceptual systems based on the modality used to receive the information, and in turn these systems interact with the language network in the brain. Overall, our brain shows a predisposition towards periodicity in the external world, which is exploited to optimally process information (VanRullen et al., 2014). As laid out in Chapter 1, the temporal resolution of the acoustic system is much higher and finer grained compared to that of the visual system (Holcombe, 2009; Moore, 1993). The results presented in this doctoral thesis support the claim that the auditory domain is more sensitive to temporal periodicity.

Chapter 2 showed that language comprehension decreases when the inherent periodic structure of language is distorted with a temporal manipulation and this decrease is common to both sign and spoken languages. However, the way that comprehension is affected differs between modalities.

Spoken language shows a clear perceptual threshold after which language comprehension is lost; this threshold has been suggested to correspond to the basic unit of temporal integration of the linguistic information over time (Poeppel, 2003). Sign language is more resilient than spoken language to temporal distortion. The simultaneous nature of language information afforded by the visual domain produces a certain degree of redundancy, which might compensate for the loss of other (temporal) information. No clear temporal integration window can be identified in the visual domain. It is possible that visual integration over time is not a fundamental part of sign language comprehension, especially at the level of single signs targeted in our study. The three fundamental sublexical features of sign language are movement, location and handshape of the sign (Herrero Blanco, 2009). Of the three, neither handshape nor location are particularly impacted by temporal manipulation and can be retrieved even in a static depiction of the sign (in the absence of movement).

Complementing this result, in Chapter 4 we saw that phase synchronization between brain oscillations and the linguistic signal, a phenomenon commonly referred to as entrainment, is a feature of the processing of spoken and signed languages. In both modalities the brain exploits the specific temporal periodicity in the signal to parse the information. Although entrainment seems to be a modality independent mechanism, the different temporal structures of sign and spoken language are reflected in the frequency profile of entrainment. Spoken languages are characterized by periodic patterns focused in delta and theta frequency bands, which overlap with specific linguistic features such as prosodic contour and syllable rate. Sign language movements instead show slow rhythmic modulations in the delta band. Another notable difference is that synchronization is much stronger in spoken languages compared to sign language. We propose two possible, interconnected, reasons for this difference. On the one hand, the two physical signals, acoustic and visual, are characterized by a different level of temporal regularity. Spoken language generally shows more periodicity compared to sign language. On the other hand, the acoustic system shows a specific predisposition to pick up on periodic structure, in contrast with the visual system. These explanations are complementary: the acoustic language signal is more periodic since its temporal structure is exploited for its perception and processing. Sign language processing uses linguistic temporal structure to a certain degree, but it is qualitatively and quantitively different compared to spoken language. It is possible that sign language processing relies on other features of the signal that are not captured by its temporal structure; in the following section we formulate some hypotheses on which features these might be and how to test them.

5.3 Limitations and future directions

This doctoral thesis is focused on the temporal structure of language. Sign language, through the use of the visual modality, makes less use of temporal structure: language processing does not rely on entrainment as much as in spoken language and language understanding is more resilient to temporal distortion of the signal. One possibility is that the spatial domain takes a more central role in sign language. Indeed, the use of space is pervasive in sign language structure (Costello, 2016; Sandler & Lillo-Martin, 2006). At the phonological level, location represents one of three basic sublexical features, and it can distinguish between minimal pairs of signs when handshape, movements and orientation are the same. Another illustration of the centrality of space are classifier constructions, where movement in the signing space is used to express motion or existence (Emmorey, 2003; Supalla, 1982). Finally, space can be used for topographical descriptions and to refer to different topics or characters in the discourse (Emmorey et al., 1995; Lillo-Martin, 1995; Quer, 2015). Linguistic theories of sign language give a clear description of how space is employed at different levels of language structure. Future research should concentrate on a better characterization of how spatial properties are exploited in sign language, and how they interact with the brain's temporal filters.

The temporal patterns of the visual signal of sign language emerge from the movements of articulators in space, and in Chapter 3 we highlighted the importance of studying each articulator individually. In this thesis we present data from a limited set of body articulators and we base our kinematic analysis on speed vector extracted from the 3D coordinates of each articulator. Facial articulators (eyes, mouth and eyebrows) might represent an interesting source of information given their specific physical characteristics: their size is smaller, which might allow faster movements, and their motion in space is much more limited compared to hands, torso or head. Exploring different measure to describe articulators' movement, such as mouth and eye aperture or body and head orientation, can provide further information. The full understanding of the physical properties of the visual signal derived from each articulator will help us to take a further step towards linking these properties with linguistic units in sign language. A spatial manipulation, instead of a temporal one, could be used to better understand the role of each articulator in sign language comprehension.

Spatially distorting different parts of the visual scene, through the use of different degrees of visual noise, could reveal the relevance of each articulator in conveying the linguistic information.

The question of how a complex multilayered signal, such as that of sign language, is processed by our cognitive system as unified linguistic input is of particular interest. A distinctive characteristic of multilayered information is that the resulting unified signal is qualitatively different from the simple sum of the properties of its parts (Partan & Marler, 1999). How do the different visual signals integrate and combine? Do they equally contribute in creating the final percept? This question can be tackled on two different levels: production and perception. In spoken language, for example, we know that the auditory signal can stand on its own most of the time, and any visual information is normally subordinate to and coupled with the auditory stream. Moreover, in multimodal spoken language the two streams of information use distinct sensory channels (auditory and visual) and therefore perception, at least in the early stages, is separate. In sign language, on the other hand, information from each articulator is conveyed through the visual modality, and more research is needed to understand how different articulators interact with each other. Clustering and dimensionality reduction analysis can help identify groups of articulators moving in sync with each other, and pinpoint dominant kinematic patterns in the unified visual signal. At the perception level, our data on language-brain entrainment can be of use in answering this question. In this thesis we show that the brain entrains with the frequencies of the right hand, but we still have to study the differential contribution of other articulators (or their combination) to entrainment and language processing.

The work and results of this PhD argue for a change in the approach to sign language research. Spoken language research rests on a broad and deep literature, and for many years researchers compared spoken and signed languages with the final aim of testing properties and theories of spoken language in the sign domain. This approach served the important purpose of establishing the validity of sign language as a natural proper language, and helped guide educational policies and the public perception of sign language. Now we are witnessing a change in approach: some constructs coming from the spoken language field may just not be adequate to describe sign language, due to the intrinsic differences between the channels involved. Looking at sign language without predictions derived from the spoken language domain will be highly informative and could lead to a better characterization of sign language.

5.4 Conclusion

In this chapter I have summarized the main findings of this doctoral work and how they contribute to the larger literature on language processing and sign language specifically. In addition to its theoretical contribution, this work also has value in terms of practical applications. A complete understanding of the physical properties of the sign language signal, and how they relate to the linguistic content, can help the development of automatic translation technology. This technology has the potential to facilitate the communication between the hearing and deaf population. In recent years, research on sign recognition and rendering has steeply advanced (Kahlon & Singh, 2021), but this task remains challenging due to the complexity of the sign language signal. On one side there is the need for algorithms able to recognize and associate the combination of movements, location and handshape to a certain meaning which then can be expressed as text or synthesised speech. Conversely, when translating from speech to sign language, the technology focuses on a realistic rendering of the sign with avatars capable of reproducing the fluid movements of natural signing. Understanding which physical properties are fundamental in the visual sign language signal to convey linguistic information could help decide which features are important to reproduce when synthesising sign language to optimize efficient and natural communication.

This doctoral work provides evidence that the visual and acoustic modality differently shape language production and perception. Sign and spoken languages both naturally evolved exploiting the constraints and strengths of the perceptual system they employ. The different nature of auditory and visual systems – visual input tends to be constant over time while auditory input changes rapidly – is reflected in two processing systems that are differentially optimized to cope with the speed of information flow (Thorne & Debener, 2014; VanRullen et al., 2014; Zoefel & VanRullen, 2015). Our findings are in line with the hypothesis proposed by Zoefel & VanRullen (2017) for general perceptual processing: periodicity in the temporal structure is fundamental for selecting and processing of stimuli in the auditory domain, but not in the visual one. This seems to be true for complex stimuli, such as language, as well. We find that sign language relies much less than spoken language on temporal structure; this is evident both at the comprehension level where sign language is more resilient to temporal distortion of the signal (Chapter 2) and processing level where entrainment is less strong than what observed in spoken language (Chapter 4). Within this framework it seems natural for sign language to make greater use of the spatial domain, compared to the time domain, to efficiently deliver information. Nevertheless, our brain processes information over time, meaning through a temporal filter. As a result, sign language represents a particularly interesting case where spatial organization interacts with the time structure of the signal.

The study of language needs to take into account modality-dependent differences in order to identify those properties of language (production and processing) that are intrinsic to language in a more abstract sense. Our results suggest that temporal structure does not play such a fundamental role in language processing, as previously hypothesized. Temporal properties are more likely to be a result of the medium used (acoustic as opposed to visual) than of the language status of the signal. Language as a communication system is highly flexible. Sign and spoken language share the same richness of semantic, syntactic and prosodic content; this extremely complex type of information structures itself in the way that best fits the modality employed to deliver and receive the linguistic information it encodes.

References

Abbs, J. H., Gracco, V. L., & Cole, K. J. (1984). Control of multimovement coordination: Sensorimotor mechansims in speech motor programming. *Journal of Motor Behavior*, *16*(2), 195–232. https://doi.org/10.1080/00222895.1984.10735318

Abercrombie, D. (2019). Elements of General Phonetics. In *Elements of General Phonetics*. https://doi.org/10.1515/9781474463775

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2009). Abnormal cortical processing of the syllable rate of speech in poor readers. *Journal of Neuroscience*, *29*(24), 7686–7693. https://doi.org/10.1523/JNEUROSCI.5242-08.2009

Alexandrou, A. M., Saarinen, T., Kujala, J., & Salmelin, R. (2020). Cortical entrainment: what we can learn from studying naturalistic speech perception. *Language, Cognition and Neuroscience*, *35*(6), 681–693. https://doi.org/10.1080/23273798.2018.1518534

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. In *Trends in Cognitive Sciences* (Vol. 4, Issue 7, pp. 267–278). Elsevier Current Trends. https://doi.org/10.1016/S1364-6613(00)01501-1

Arnheim, R., & McNeill, D. (1994). Hand and Mind: What Gestures Reveal about Thought. *Leonardo*, *27*(4), 358. https://doi.org/10.2307/1576015

Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, *147*, 100–105. https://doi.org/10.1016/j.cognition.2015.11.013

Bartlett, A. M., Ovaysikia, S., Logothetis, N. K., & Hoffman, K. L. (2011). Saccades during object viewing modulate oscillatory phase in the superior temporal sulcus. *Journal of Neuroscience*, *31*(50), 18423–18432. https://doi.org/10.1523/JNEUROSCI.4102-11.2011

Bavelier, D., Tomann, A., Hutton, C., Mitchell, T., Corina, D., Liu, G., & Neville, H. (2000). Visual attention to the periphery is enhanced in congenitally deaf individuals. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *20*(17). https://doi.org/10.1523/jneurosci.20-17-j0001.2000

Bellugi, U., & Fischer, S. (1972). A comparison of sign language and spoken language. *Cognition*, *1*(2–3), 173–200. https://doi.org/10.1016/0010-0277(72)90018-2

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer [Computer program] Version 6.1. 16. Online at http://www. praat. org.*

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Bourguignon, M., De Tiège, X., De Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., Goldman, S., Hari, R., & Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping*, *34*(2), 314–326. https://doi.org/10.1002/hbm.21442

Brentari, D. (1998). *A prosodic model of sign language phonology.* Mit Press.

Brookshire, G., Lu, J., Nusbaum, H. C., Goldin-Meadow, S., & Casasanto, D. (2017). Visual cortex entrains to sign language. *Proceedings of the National Academy of Sciences*, *114*(24), 6352–6357. https://doi.org/10.1073/pnas.1620350114

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

Carreiras, M., Gutiérrez-Sigut, E., Baquero, S., & Corina, D. (2008). Lexical processing in Spanish Sign Language (LSE). *Journal of Memory and Language*, *58*(1), 100–122. https://doi.org/10.1016/j.jml.2007.05.004

Caselli, N. K., Emmorey, K., & Cohen-Goldberg, A. M. (2021). The signed mental lexicon: Effects of phonological neighborhood density, iconicity, and childhood language experience. *Journal of Memory and Language*, *121*, 104282. https://doi.org/10.1016/J.JML.2021.104282

Cormier, K., Smith, S., & Zwets, M. (2013). Framing constructed action in British Sign Language narratives. *Journal of Pragmatics*, *55*, 119–139. https://doi.org/10.1016/j.pragma.2013.06.002

Costello, B. D. N. (2016). Language and modality: Effects of the use of space in the agreement system of lengua de signos española (Spanish Sign Language). In *Sign Language and Linguistics (Online)* (Vol. 19, Issue 2).

Coulter, G. (1982). On the nature of ASL as a monosyllabic language. *56th Annual Meeting of the Linguistic Society of America*.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Cummins, F. (2012). Oscillators and syllables: A cautionary note. *Frontiers in Psychology*, *3*(OCT). https://doi.org/10.3389/fpsyg.2012.00364

Cutini, S., Szűcs, D., Mead, N., Huss, M., & Goswami, U. (2016). Atypical right hemisphere response to slow temporal modulations in children with developmental dyslexia. *NeuroImage*, *143*, 40–49. https://doi.org/10.1016/j.neuroimage.2016.08.012

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*(4), 385–400. https://doi.org/10.1016/0749-596X(86)90033-1

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, *71*, 268–283. https://doi.org/10.1016/j.wocn.2018.09.007

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Wens, V., De Tiège, X., & Bourguignon, M. (2019). Comparing the potential of MEG and EEG to uncover brain tracking of speech temporal envelope. *NeuroImage*, *184*, 201–213. https://doi.org/10.1016/j.neuroimage.2018.09.006

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, *25*(19). https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in Human Neuroscience*, *11*(September), 1–9. https://doi.org/10.3389/fnhum.2017.00481

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Doelling, K., Arnal, L., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85*(15). https://doi.org/10.1016/j.neuroimage.2013.06.035.Acoustic

Drullman, R. (2019). The Significance of Temporal Modulation Frequencies for Speech Intelligibility. In *Listening to Speech* (pp. 39–47). Psychology Press. https://doi.org/10.4324/9780203933107-3

Dye, M. W. (2006). *Phonological priming in British Sign Language Development of Temporal Visual Selective Attention in Deaf Children View project Project DyAdd View project*.

Efrón, D. (1941). *Gesture and Environment* (King's crown Press (ed.)).

Emmorey, K. (2003). Perspectives on classifier constructions in sign languages. In *Perspectives on Classifier Constructions in Sign Languages*. https://doi.org/10.4324/9781410607447

Emmorey, K. (2021). New Perspectives on the Neurobiology of Sign Languages. *Frontiers in Communication*, *6*(December), 1–20. https://doi.org/10.3389/fcomm.2021.748430

Emmorey, K., Corina, D., & Bellugi, U. (1995). Differential processing of topographic and referential functions of space. In *Language, Gesture, and Space* (pp. 43–62). https://psycnet.apa.org/record/1995-97525-002

Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, *2*(JUN), 130. https://doi.org/10.3389/fpsyg.2011.00130

Ghitza, O. (2013). The theta-syllable: A unit of speech information defined by cortical function. *Frontiers in Psychology*, *4*(MAR). https://doi.org/10.3389/fpsyg.2013.00138

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, *66*(1–2), 113–126. https://doi.org/10.1159/000208934

Giraud, A. L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. *Neuron*, *56*(6), 1127–1134. https://doi.org/10.1016/j.neuron.2007.09.038

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. In *Nature Neuroscience* (Vol. 15, Issue 4, pp. 511–517). https://doi.org/10.1038/nn.3063

Goswami, U., & Leong, V. (2013). Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology*, *4*(1). https://doi.org/10.1515/lp-2013-0004

Greenberg, S., & Arai, T. (2004). *What are the Essential Cues for Understanding Spoken Language?*

Grosjean, F. (1977a). The perception of rate in spoken and sign languages. *Perception & Psychophysics*, *22*(4), 408–413. https://doi.org/10.3758/BF03199708

Grosjean, F. (1977b). The perception of rate in spoken and sign languages. *Perception & Psychophysics*, *22*(4), 408–413. https://doi.org/10.3758/BF03199708

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain. *PLoS Biology*, *11*(12), e1001752. https://doi.org/10.1371/journal.pbio.1001752

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. *Current Biology*, *28*(9), 1453-1459.e3. https://doi.org/10.1016/j.cub.2018.03.044

Herrero Blanco, A. (2009). *Gramática didáctica de la lengua de signos española (LSE).* Sm.

Hildebrandt, U., & Corina, D. (2002). Phonological similarity in American Sign Language. In *Language and Cognitive Processes* (Vol. 17, Issue 6, pp. 593–612). https://doi.org/10.1080/01690960143000371

Hoffman, K. L., Dragan, M. C., Leonard, T. K., Micheli, C., Montefusco-Siegmund, R., & Valiante, T. A. (2013). Saccades during visual exploration align hippocampal 3–8 Hz rhythms in human and non-human primates. *Frontiers in Systems Neuroscience*, *7*, 43. https://doi.org/10.3389/fnsys.2013.00043

Holcombe, A. O. (2009). Seeing slow and seeing fast: two limits on perception. *Trends in Cognitive Sciences*, *13*(5), 216–221. https://doi.org/10.1016/j.tics.2009.02.005

Howard, M. F., & Poeppel, D. (2012). The neuromagnetic response to spoken sentences: Co-modulation of theta band amplitude and phase. *NeuroImage*, *60*(4), 2118–2127. https://doi.org/10.1016/j.neuroimage.2012.02.028

Huggins, A. W. F. (1975). Temporally segmented speech. *Perception & Psychophysics*, *18*(2), 149–157. https://doi.org/10.3758/BF03204103

Hwang, S.-O. K. (2011). Windows into sensory integration and rates in language processing: Insights from signed and spoken languages. In *Dissertation Abstracts International Section A: Humanities and Social Sciences* (Vol. 73, Issues 6-A).

Ishida, M. (2021). Perceptual restoration of locally time-reversed speech: Non-native listeners' performance in their L2 vs. L1. *Attention, Perception, and Psychophysics*, *83*(6), 2675–2693. https://doi.org/10.3758/s13414-021-02258-5

Ishida, M., Arai, T., & Kashino, M. (2018). Perceptual restoration of temporally distorted speech in L1 vs. L2: Local time reversal and modulation filtering. *Frontiers in Psychology*, *9*(SEP), 1–16. https://doi.org/10.3389/fpsyg.2018.01749

Ito, J., Maldonado, P., Singer, W., & Grün, S. (2011). Saccade-related modulations of neuronal excitability support synchrony of visually elicited spikes. In *Cerebral Cortex* (Vol. 21, Issue 11, pp. 2482–2497). https://doi.org/10.1093/cercor/bhr020

Janzen, T. (2004). Space rotation, perspective shift, and verb morphology in ASL. *Cognitive Linguistics*, *15*(2), 149–174. https://doi.org/10.1515/cogl.2004.006

Jin, S.-H., & Nelson, P. B. (2010). Interrupted speech perception: The effects of hearing sensitivity and frequency resolution. *The Journal of the Acoustical Society of America*, *128*(2), 881–889. https://doi.org/10.1121/1.3458851

Kahlon, N. K., & Singh, W. (2021). Machine translation from text to sign language: a systematic review. In *Universal Access in the Information Society* (Issue 0123456789). Springer Berlin Heidelberg. https://doi.org/10.1007/s10209-021-00823-1

Keitel, A., Ince, R. A. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *NeuroImage*, *147*, 32–42. https://doi.org/10.1016/j.neuroimage.2016.11.062

Kendon, A. (2015). Gesture: Visible action as utterance. In *Gesture: Visible Action as Utterance*. Cambridge University Press. https://doi.org/10.5860/choice.42-5687

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Kiss, M., Cristescu, T., Fink, M., & Wittmann, M. (2008). Auditory language comprehension of temporally reversed speech signals in native and non-native speakers. *Acta Neurobiologiae Experimentalis*, *68*(2), 204–213.

Klima, E. S., & Bellugi, U. (1979). *The signs of language.* Harvard University Press.

Kösem, A., & van Wassenhove, V. (2017). Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, *32*(5), 536–544. https://doi.org/10.1080/23273798.2016.1238495

Lakatos, P., Gross, J., & Thut, G. (2019). Review A new unifying account of the roles of neuronal entrainment. *Current Biology*, *29*(18), 1–16. https://doi.org/10.1016/j.cub.2019.07.075

Lehongre, K., Ramus, F., Villiermet, N., Schwartz, D., & Giraud, A. L. (2011). Altered Low-Gamma Sampling in Auditory Cortex Accounts for the Three Main Facets of Dyslexia. *Neuron*, *72*(6), 1080–1090. https://doi.org/10.1016/J.NEURON.2011.11.002

Leonard, M. K., Ramirez, N. F., Torres, C., Travis, K. E., Hatrak, M., Mayberry, R. I., & Halgren, E. (2012). Signed words in the congenitally deaf evoke typical late lexicosemantic responses with no early visual responses in left superior temporal cortex. *Journal of Neuroscience*, *32*(28), 9700–9705. https://doi.org/10.1523/JNEUROSCI.1002-12.2012

Levänen, S., Uutela, K., Salenius, S., & Hari, R. (2001). Cortical representation of sign language: Comparison of deaf signers and hearing non-signers. *Cerebral Cortex*, *11*(6), 506–512. https://doi.org/10.1093/cercor/11.6.506

Lieberman, A. M., & Borovsky, A. (2020). Lexical Recognition in Deaf Children Learning American Sign Language: Activation of Semantic and Phonological Features of Signs.

Language Learning, *70*(4), 935–973. https://doi.org/10.1111/lang.12409

Lillo-Martin, D. (1995). The Point of View Predicate in American Sign Language. In K. Emmorey & J. S. Reilly (Eds.), *Language, Gesture, and Space* (pp. 155–170). Psychology Press. https://doi.org/10.4324/9780203773413-14

Lindblad, P., Karlsson, S., & Heller, E. (1991). Mandibular movements in speech phrases - a syllabic quasiregular continuous oscillation. *Logopedics Phoniatrics Vocology*, *16*(1–2), 36–42. https://doi.org/10.3109/14015439109099172

Lizarazu, M., Carreiras, M., Bourguignon, M., Zarraga, A., & Molinaro, N. (2021). Language Proficiency Entails Tuning Cortical Activity to Second Language Speech. *Cerebral Cortex*, *31*(8), 3820–3831. https://doi.org/10.1093/cercor/bhab051

Annals of the New York Academy of Sciences, *1453*(1), 140–152. https://doi.org/10.1111/nyas.14099

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS Biology*, *8*(8), 25–26. https://doi.org/10.1371/journal.pbio.1000445

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

MacSweeney, M., Capek, C. M., Campbell, R., & Woll, B. (2008). The signing brain: the neurobiology of sign language. *Trends in Cognitive Sciences*, *12*(11), 432–440. https://doi.org/10.1016/j.tics.2008.07.010

Malaia, E. A., Borneman, S. C., Krebs, J., & Wilbur, R. B. (2021). Low-Frequency Entrainment to Visual Motion Underlies Sign Language Comprehension. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 2456–2463. https://doi.org/10.1109/TNSRE.2021.3127724

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Matsuo, I., Ueda, K., & Nakajima, Y. (2020). Intelligibility of chimeric locally time-reversed speech. *The Journal of the Acoustical Society of America*, *147*(6), EL523–EL528. https://doi.org/10.1121/10.0001414

Mayberry, R. I., & Fischer, S. D. (1989). Looking through phonological shape to lexical meaning: The bottleneck of non-native sign language processing. *Memory & Cognition*, *17*(6), 740–754. https://doi.org/10.3758/BF03202635

Mayberry, R. I., & Witcher, P. (2005). What Age of Acquisition Effects Reveal about the Nature of Phonological Processing. *CRL Technical Reports*, *17*(3), 3–9.

McNeill, D. (2013). Gesture and Thought. In *Gesture and Thought*. https://doi.org/10.7208/chicago/9780226514642.001.0001

Meade, G., Lee, B., Midgley, K. J., Holcomb, P. J., & Emmorey, K. (2018). Phonological and semantic priming in american sign language: N300 and N400 effects. *Language, Cognition and Neuroscience*, *33*(9), 1092–1106. https://doi.org/10.1080/23273798.2018.1446543

Meier, R. P. (1991). Language acquisition by deaf children. In *American Scientist*.

Meier, R. P. (2002). Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In *Modality and structure in signed and spoken languages* (pp. 1–25).

Mersad, K., Goyet, L., & Nazzi, T. (2011). Cross-linguistic differences in early word form segmentation: a rhythmic-based account. *Journal of Portuguese Linguistics*, *10*(1), 37. https://doi.org/10.5334/jpl.100

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. In *Nature* (Vol. 485, Issue 7397, pp. 233–236). NIH Public Access. https://doi.org/10.1038/nature11020

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, *48*(7), 2609–2621. https://doi.org/10.1111/ejn.13748

Meyer, L., & Gumbert, M. (2018). Synchronization of electrophysiological responses with speech benefits syntactic information processing. *Journal of Cognitive Neuroscience*, *30*(8), 1066–1074. https://doi.org/10.1162/jocn_a_01236

Miller, G. A., & Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *Journal of the Acoustical Society of America*, *22*(2), 167–173. https://doi.org/10.1121/1.1906584

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, *48*(7), 2642–2650. https://doi.org/10.1111/ejn.13811

Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., & Carreiras, M. (2016). Out-of-synchrony speech entrainment in developmental dyslexia. *Human Brain Mapping*, *37*(8), 2767–2783. https://doi.org/10.1002/hbm.23206

Moore, B. C. J. (1993). Temporal Analysis in Normal and Impaired Hearing. *Annals of the New York Academy of Sciences*, *682*(1), 119–136. https://doi.org/10.1111/j.1749-

6632.1993.tb22964.x

Nacar Garcia, L., Guerrero-Mosquera, C., Colomer, M., & Sebastian-Galles, N. (2018). Evoked and oscillatory EEG activity differentiates language discrimination in young monolingual and bilingual infants. *Scientific Reports 2018 8:1*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-20824-0

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language Discrimination by Newborns: Toward an Understanding of the Role of Rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(3), 756–766. https://doi.org/10.1037/0096-1523.24.3.756

Nelson, P. B., & Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *115*(5), 2286–2294. https://doi.org/10.1121/1.1703538

Nespor, M., & Sandler, W. (1999). Prosody in Israeli sign language. *Language and Speech*, *42*(2–3), 143–176. https://doi.org/10.1177/00238309990420020201

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-Timed vs . Syllable-Timed Languages. In *The Blackwell Companion to Phonology* (pp. 1–13). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444335262.wbctp0048

Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1207/s15516709cog1401_2

Newport, E. L., & Meier, R. P. (2018). The Acquisition of American Sign Language. In *The Crosslinguistic Study of Language Acquisition* (pp. 881–938). Psychology Press. https://doi.org/10.4324/9781315802541-12

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/CERCOR/BHT355

Obleser, J., & Kayser, C. (2019). Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences*, *23*(11), 913–926. https://doi.org/10.1016/j.tics.2019.08.004

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1651). https://doi.org/10.1098/rstb.2013.0296

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, *16*(8), 1–26.

https://doi.org/10.1371/journal.pbio.2006558

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *ELife*, *5*(MAY2016). https://doi.org/10.7554/eLife.14521

Partan, S., & Marler, P. (1999). Communication goes multimodal. In *Science* (Vol. 283, Issue 5406, pp. 1272–1273). American Association for the Advancement of Science. https://doi.org/10.1126/science.283.5406.1272

Pastureau, R. (2022). *Krajjat: Kinect Realignment Algorithm for Joint Jumps And Twitches (Version 1.6) [Computer software]*.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Peña, M., & Melloni, L. (2012). Brain oscillations during spoken sentence processing. *Journal of Cognitive Neuroscience*, *24*(5), 1149–1164. https://doi.org/10.1162/jocn_a_00144

Peters, R. W., & Glasberq, B. R. (1993). Detection of temporal gaps in sinusoids: Effects of frequency and level. *Journal of the Acoustical Society of America*, *93*(3), 1563–1570. https://doi.org/10.1121/1.406815

Petitto, L. A., & Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, *251*(5000), 1493–1496. https://doi.org/10.1126/science.2006424

Pfau, R., Steinback, M., & Woll, B. (2012). Sign language: An international handbook. In *Phonetics* (pp. 4–20).

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time." *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-020-0304-4

Pouw, W., & Dixon, J. a. (2019). *Quantifying gesture-speech synchrony*. *6*, 75–80. https://doi.org/10.17619/UNIPB/1-815

Proksch, J., & Bavelier, D. (2002). Changes in the spatial distribution of visual attention after early deafness. *Journal of Cognitive Neuroscience*, *14*(5), 687–701. https://doi.org/10.1162/08989290260138591

Puce, A., & Perrett, D. (2003). Electrophysiology and brain imaging of biological motion. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 358, Issue 1431, pp. 435–445). https://doi.org/10.1098/rstb.2002.1221

Quer, J. (2015). Context Shift and Indexical Variables in Sign Languages. *Semantics and Linguistic Theory*, *15*, 152. https://doi.org/10.3765/salt.v15i0.2923

Ramus, F., Nespor, M., & Mehler, J. (2000). Correlates of linguistic rhythm in the speech signal. *Cognition*, *75*(1), 265–292. https://doi.org/10.1016/s0010-0277(00)00101-3

Reilly, J. S., Mcintire, M., & Bellugi, U. (1990). The acquisition of conditionals in American Sign Language: Grammaticized facial expressions. *Applied Psycholinguistics*, *11*(4), 369–392. https://doi.org/10.1017/S0142716400009632

Riely, R. R., & Smith, A. (2003). Speech movements do not scale by orofacial structure size. *Journal of Applied Physiology*, *94*(6), 2119–2126. https://doi.org/10.1152/japplphysiol.00502.2002

Rivolta, C. L., Costello, B., & Carreiras, M. (2021). Language modality and temporal structure impact processing: Sign and speech have different windows of integration. *Journal of Memory and Language*, *121*, 104283. https://doi.org/10.1016/j.jml.2021.104283

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, *398*(6730), 760. https://doi.org/10.1038/19652

Saija, J. D., Akyürek, E. G., Andringa, T. C., & Başkent, D. (2014). Perceptual restoration of degraded speech is preserved with advancing age. *JARO - Journal of the Association for Research in Otolaryngology*, *15*(1), 139–148. https://doi.org/10.1007/s10162-013-0422-z

Samuel, A. G. (1991). Perceptual Degradation Due to Signal Alternation: Implications for Auditory Pattern Processing. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 392–403. https://doi.org/10.1037/0096-1523.17.2.392

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*(November 2019), 104070. https://doi.org/10.1016/j.jml.2019.104070

Sandler, W. (2011). *Phonological representation of the sign: Linearity and nonlinearity in American Sign Language: Vol. (Vol. 32).* Gruyter., Walter de.

Sandler, Wendy. (2018). The body as evidence for the nature of language. *Frontiers in Psychology*, *9*(OCT), 1–21. https://doi.org/10.3389/fpsyg.2018.01782

Sandler, Wendy, & Lillo-Martin, D. (2006). Sign language and linguistic universals. In *Sign Language and Linguistic Universals*. Cambridge University Press. https://doi.org/10.1017/CBO9781139163910

Shafiro, V., Sheft, S., Kuvadia, S., & Gygi, B. (2015). Environmental sound training in cochlear implant users. *Journal of Speech, Language, and Hearing Research*, *58*(2), 509–519. https://doi.org/10.1044/2015_JSLHR-H-14-0312

Shafiro, V., Sheft, S., & Risley, R. (2016). The intelligibility of interrupted and temporally altered speech: Effects of context, age, and hearing loss. *The Journal of the Acoustical Society of America*, *139*(1), 455–465. https://doi.org/10.1121/1.4939891

Steffen, A., & Werani, A. (1994). An experiment on temporal processing in language perception. *Sprechwissenschaft Und Psycholinguistik*, *6*, 189–205.

Stilp, C. E., Kiefte, M., Alexander, J. M., & Kluender, K. R. (2010). Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentencesa). *The Journal of the Acoustical Society of America*, *128*(4), 2112–2126. https://doi.org/10.1121/1.3483719

Stokoe, W. C., & Marschark, M. (2005). Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, *10*(1), 3–37. https://doi.org/10.1093/deafed/eni001

Supalla, T. E. D. R. (1982). Structure and Acquisition of Verbs of Motion and Location in American Sign Language. *ProQuest Dissertations and Theses*, *November 1982*, 149. https://psycnet.apa.org/record/1983-70283-001

Teng, X., Cogan, G. B., & Poeppel, D. (2019). Speech fine structure contains critical temporal cues to support speech segmentation. *NeuroImage*, *202*. https://doi.org/10.1016/j.neuroimage.2019.116152

Thomson, J. M., & Goswami, U. (2008). Rhythmic processing in children with developmental dyslexia: Auditory and motor rhythms link to reading and spelling. *Journal of Physiology Paris*, *102*(1–3), 120–129. https://doi.org/10.1016/j.jphysparis.2008.03.007

Thorne, J. D., & Debener, S. (2014). Look now and hear what's coming: On the functional role of cross-modal phase reset. In *Hearing Research* (Vol. 307, pp. 144–152). Elsevier. https://doi.org/10.1016/j.heares.2013.07.002

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *146*(10).

Ueda, K., Nakajima, Y., Ellermeier, W., & Kattner, F. (2017). Intelligibility of locally time-reversed speech: A multilingual comparison. *Scientific Reports*, *7*(1), 1–8. https://doi.org/10.1038/s41598-017-01831-z

Vanden Bosch der Nederlanden, C. M., Joanisse, M. F., Grahn, J. a., Snijders, T. M., & Schoffelen, J. M. (2022). Familiarity modulates neural tracking of sung and spoken utterances. *NeuroImage*, *252*(March), 119049. https://doi.org/10.1016/j.neuroimage.2022.119049

Vander Ghinst, M., Bourguignon, M., Op de Beeck, M., Wens, V., Marty, B., Hassid, S., Choufani, G., Jousmäki, V., Hari, R., Van Bogaert, P., Goldman, S., & De Tiège, X. (2016). Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. *Journal of Neuroscience*, *36*(5), 1596–1606. https://doi.org/10.1523/JNEUROSCI.1730-15.2016

VanRullen, R., Zoefel, B., & Ilhan, B. (2014). On the cyclic nature of perception in vision versus audition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1641). https://doi.org/10.1098/rstb.2013.0214

Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. *Citation: The Journal of the Acoustical Society of America*, *90*, 858. https://doi.org/10.1121/1.401953

Villameriel, S., Costello, B., Dias, P., Giezen, M., & Carreiras, M. (2019). Language modality shapes the dynamics of word and sign recognition. *Cognition*, *191*, 103979. https://doi.org/10.1016/j.cognition.2019.05.016

Wagner, P., Malisz, Z., & Kopp, S. (2014a). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wagner, P., Malisz, Z., & Kopp, S. (2014b). Gesture and speech in interaction: An overview. In *Speech Communication* (Vol. 57, pp. 209–232). North-Holland. https://doi.org/10.1016/j.specom.2013.09.008

Walsh, B., & Smith, A. (2002). Articulatory movements in adolescents: Evidence for protracted development of speech motor control processes. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1119–1133. https://doi.org/10.1044/1092-4388(2002/090)

Wilbur, R. B. (2009). Effects of Varying Rate of Signing on ASL Manual Signs and Nonmanual Markers. *Language and Speech*, *52*(2–3), 245–285. https://doi.org/10.1177/0023830909103174

Zoefel, B., & VanRullen, R. (2015). Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. *Journal of Neuroscience*, *35*(5), 1954–1964. https://doi.org/10.1523/JNEUROSCI.3484-14.2015

Zoefel, B., & VanRullen, R. (2017). Oscillatory mechanisms of stimulus processing and selection in the visual and auditory systems: State-of-the-art, speculations and suggestions. *Frontiers in Neuroscience*, *11*(MAY), 1–13. https://doi.org/10.3389/fnins.2017.00296

Appendices

Appendix 1

Table A1: Summary of results of post-hoc t-tests (corrected using the Bonferroni method) comparing intelligibility between paired reversal windows for the Spanish and LSE stimuli. Post-hoc statistics are based on fitted data from the linear mixed model.

Reversal window	**Reversal window duration (ms)**		**Comparison of paired reversal windows across tasks**	
	Spanish	**LSE**	***t***	***p***
1	0	0	-5.30	< .001
2	40	133	-12.11	< .001
3	55	199	0.11	1.00
4	70	266	19.34	< .001
5	85	333	24.43	< .001
6	100	399	24.91	< .001

Table A2: Summary of results of post-hoc t-tests (corrected using the Bonferroni method) comparing intelligibility between paired reversal windows for LSE and visual non-linguistic stimuli. Post-hoc statistics are based on fitted data from the linear mixed model.

Reversal window	**Reversal window duration (ms)**		**Comparison of paired reversal windows across tasks**	
	LSE	**Non-ling**	***t***	***p***
1	0	0	-0.39	1.00
2	133	133	-0.57	1.00
3	199	199	5.59	< .001
4	266	266	9.41	< .001
5	333	333	11.36	< .001
6	399	399	11.87	< .001

Appendix 2

Table A3: Table of body points tracked by the custom-made Kinect system. The list includes 21 points corresponding to 21 body joints, and 68 points corresponding to 7 face features.

Body	
Point	***Label***
1	Head
2	Neck
3	SpineShoulder
4	SpineMid
5	SpineBase
6	ShoulderLeft
7	ElbowLeft
8	WristLeft
9	HandLeft
10	ShoulderRight
11	ElbowRight
12	WristRight
13	HandRight
14	HipLeft
15	KneeLeft
16	AnkleLeft
17	FootLeft
18	HipRight
19	KneeRight
20	AnkleRight
21	FootRight

Face	
Point	***Label***
0 - 17	Chin
18 - 22	Eyebrow Left
23 - 27	Eyebrow Right
28 - 36	Nose
37 - 42	Eye Left
43 - 48	Eye Right
49 - 69	Mouth

Table A4: Descriptive statistics in Hertz (mean, standard deviations, minimum and maximum) of variable frame rates of Kinect recordings. The statistics are presented separately for each language.

	Mean	***Standard Dev***	***Minimum***	***Maximum***
LSE	23.94	1.93	7.43	31.37
RSL	20.34	2.34	9.12	27.10
Spanish	21.37	2.23	1.41	27.90
Russian	20.41	2.09	8.54	26.40

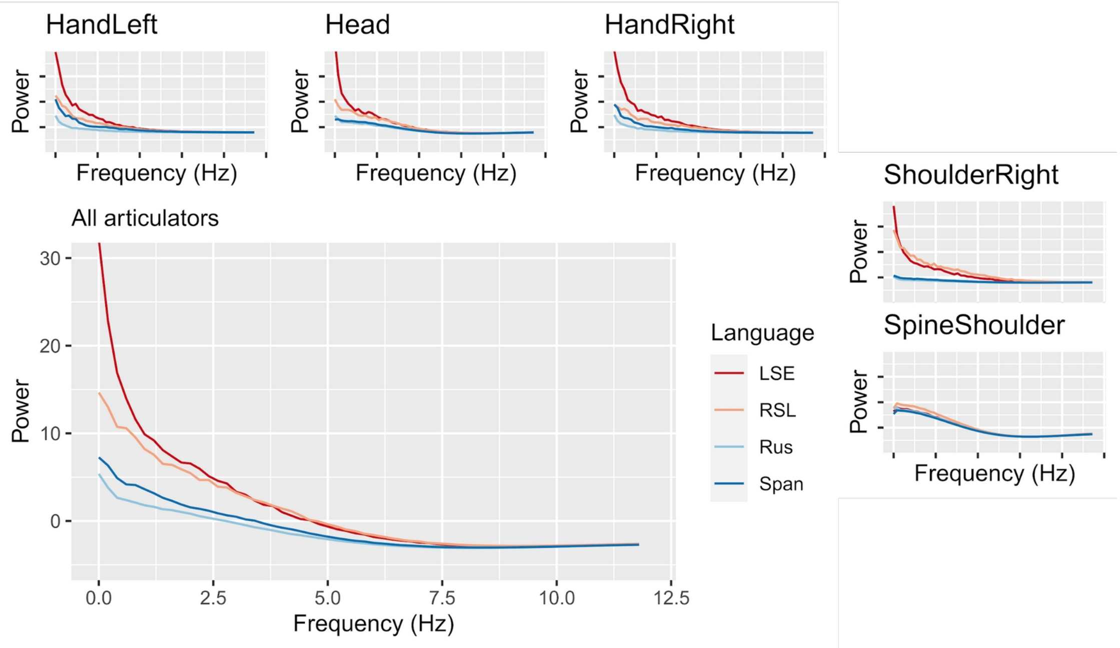

Figure A1: Time-frequency plots across languages. The central plot represents the power spectrum of all the articulators of interest, with smaller plots for each articulator.

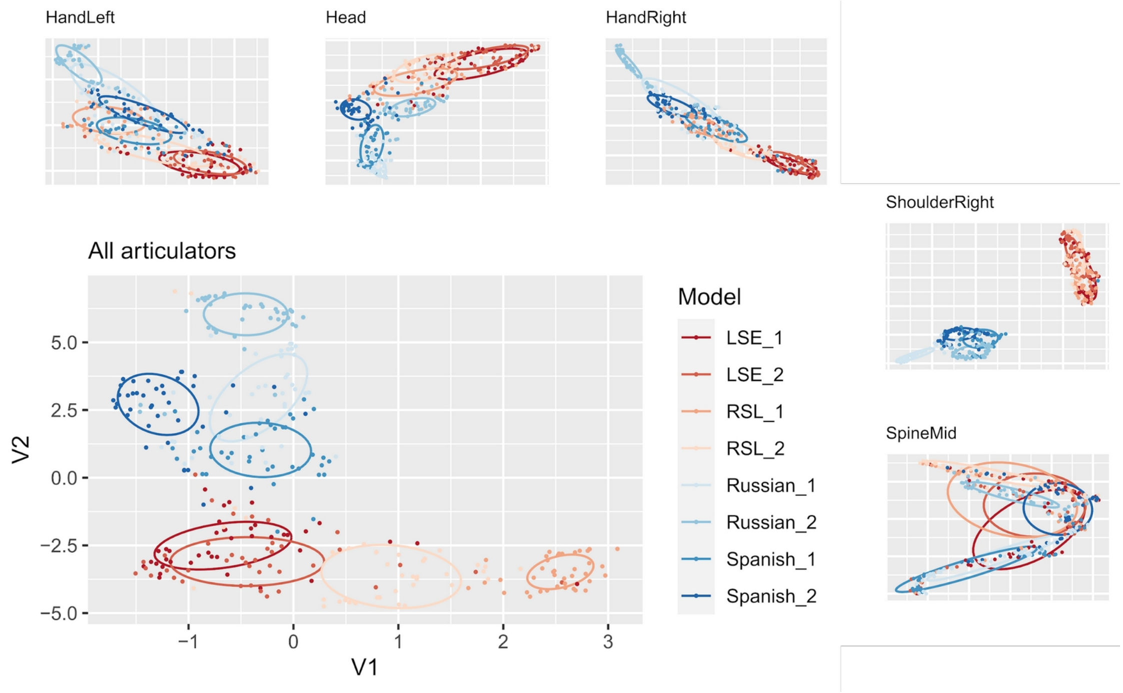

Figure A2: UMAP clusters across all models. Each dot represents a video of our dataset, and the

classification is based on the time-frequency profile derived from each video. The larger plot represent the cluster based on all the articulators of interest, with smaller plots for each articulator.

Appendix 3

In section 4.4.2.2 we report the results of the interaction analysis between language knowledge and language modality in control participants. In the theta band, due to the lack of an interaction, we can only look at main effects. These main effects are not meaningful: the lack of knowledge of LSE in this group gives rise to an imbalance of known and unknown languages in the comparisons (Spanish+LSE versus Russian+RSL; Spanish+Russian versus LSE+RSL). For the sake of completeness, these main effect contrasts are reported here. Following the same logic, the simple contrast between Spanish and LSE (in the delta band) confounds both modality and language knowledge for this group, and therefore we report it here.

Language knowledge

As noted above, this contrast does not actually reflect language knowledge for the control group, who did not know LSE.

Theta: The comparison between Spanish languages (Spanish plus LSE) and Russian languages (Russian plus RSL) shows a right temporal cluster indicating more coherence to the latter with respect to the former (p = 0.006).

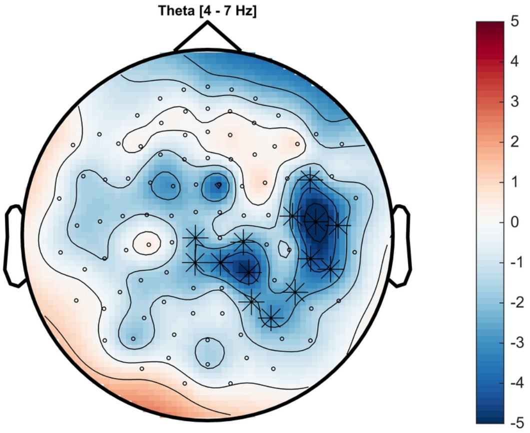

Figure A3: Plots showing the difference in coherence between Spanish and Russian languages in theta frequency band in controls (n = 14).

Language modality

Theta: The comparison between spoken and signed languages shows a bilateral temporal cluster indicating more coherence in spoken languages compared to signed languages (p = 0.002).

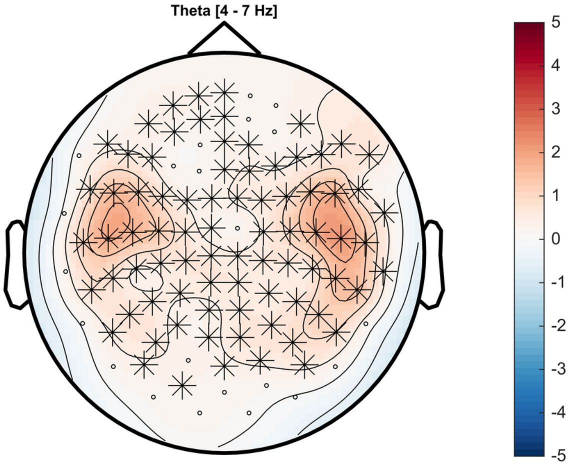

Figure A4: Plots showing the difference in coherence between spoken and signed languages in theta frequency band in controls (n = 14).

Delta: The comparison between Spanish and LSE shows two temporal clusters indicating more coherence in Spanish compared to LSE (p = 0.006, p = 0.012).

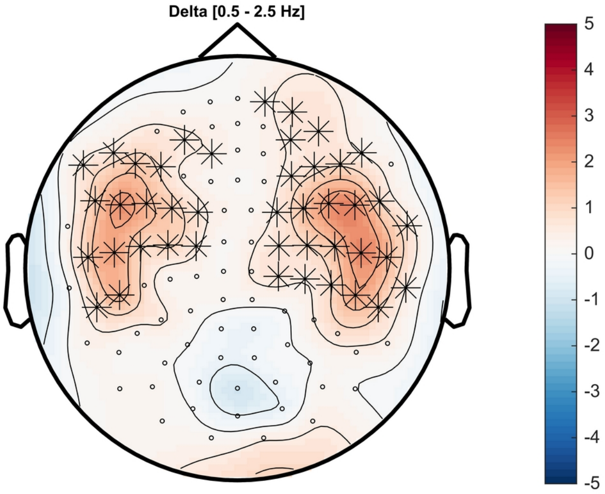

Figure A5: Plots showing the difference in coherence between Spanish and LSE in delta frequency band in controls (n = 14).