Noname manuscript No.
(will be inserted by the editor)

PRAUTOCAL Corpus:

A corpus for the study of Down syndrome prosodic aspects.

David Escudero-Mancebo · Mario Corrales-Astorgano · Valentín Cardeñoso-Payo · Lourdes Aguilar · César González-Ferreras · Pastora Martínez-Castilla · Valle Flores-Lucas

Received: date / Accepted: date

Abstract Oral productions of speakers with Down syndrome exhibit special characteristics that have been the target of study for decades. In spite of this attention, the availability of rich resources for its analysis is still scarce. In this paper, we present the definition and compiling procedure of a corpus of semi-controlled oral productions of speakers with Down syndrome that aims to allow the analysis of how these speakers with these speakers produce functional and linguistic aspects of speech. The PRAUTOCAL corpus has been recorded while using a video game for training oral competences. Utterances are related to well defined communicative tasks recorded by both speakers with Down syndrome and typically developing speakers. We present the procedure for human experts to evaluate the recordings and the transcription criteria followed for enriching the utterances of the corpus. PRAUTOCAL permits the analysis of the clear contrast in voice and speech between individuals with Down syndrome and typically developing speakers, taking into account the high heterogeneity of the speech problems characteristic of the syndrome. This material allows the analysis of the speech problems in Down syndrome, with applications to the generation of knowledge that could be used in future works for therapists to prepare specific training or enriching diagnosis regarding possible speech and language disorders.

The activities of Down syndrome speech analysis continue (1/2018-12/2020) in the project funded by the Ministerio de Ciencia, Innovación y Universidades and the European Regional Development Fund FEDER (TIN2017-88858-C2-1-R) and in the project funded by the Junta de Castilla y León (VA050G18). Part of this work was funded by the BBVA Foundation (2015-2017) in the framework of the project PRADIA: Pragmatics and prosody: the graphic adventure game.

D. Escudero-Mancebo
Departamento de Informática. Grupo ECA-SIMM.
Universidad de Valladolid.
Tel.: +34-983-185647
Fax: +34-983-423671
E-mail: descuder@infor.uva.es

Keywords Down syndrome, corpus collection with video games, evaluation of prosodic quality

1 Introduction

In all circumstances, the process of collecting a spoken corpus is difficult, time consuming and expensive. However, compiling a corpus of controlled samples of speech of individuals with Down syndrome (DS) poses extra challenges, due to the special characteristics of the informants: among others, they can suffer developmental language problems (Martin et al., 2009), attention impairments (Martínez et al., 2011) and/or short-term memory problems (Chapman and Hesketh, 2001). In particular, the latter two make it difficult for speakers with the syndrome to follow instructions and focus on speech tasks during medium length and long recording sessions. The lack of a reference corpus of comparable speech of typically developing (TD) (i.e. without disabilities) individuals also hampers the study of the speech and language characteristics of this population. Nevertheless, there is a growing need for training data to develop automated tools -specifically designed for individuals with special needs- that use spoken language technology: automatic speech recognition, computer-aided speech therapy tools, learning tools. In all these cases, the underlying machine learning models and the use of deep learning technologies requires large quantities of data. In this paper, we show how, as an alternative to classical recording corpus procedures, using a video game is a strategy that permits speakers' fatigue and loss of attention to be reduced. Over different recording campaigns, we have collected the corpus PRAUTOCAL described in the following sections.

PRAUTOCAL is a corpus of northern/central peninsular Spanish speakers with DS that has an important volume of transcribed and annotated recordings, as well as quality assessments and references of TD users for comparison. We expect the availability of PRAUTOCAL will favour the analysis of specific aspects of the speech of individuals with DS, thus generating a body of knowledge that will allow therapists to arrange specific speech and language training in future applications and/or enrich the diagnosis of possible speech and language disorders.

Our video game was designed to train oral competences related to prosody and pragmatics in a realistic context for the Spanish language. It includes pre-established oral production activities that take into account such specific linguistic and communicative functions as learning objectives (see section 3.1). This fact contributes to the creation of a corpus with utterances classified in terms of prosodic function, production mode (read, elicited, imitation or spontaneous speech) and linguistic politeness (see section 3.2). The fundamental idea for promoting the efficient training of players during game sessions is that users must express themselves correctly to keep on playing. Every time the user does not utter appropriately, he or she must repeat the activity. This fact permits PRAUTOCAL to be annotated in terms of quality. In previous works, we set the quality assessment information with off-line evaluations (Corrales-Astorgano et al., 2019). In this paper, we add subjective evaluations following a rubric that includes information about fluency, intonation curve, speech rate, intelligibility and correctness and word omissions. Subjective judgments of quality are a valuable source of information, especially when they come from speech therapists. In PRAUTOCAL, this information is complemented with a reference concerning goodness of pronunciation provided by recordings of the same sentences uttered by TD speakers.

To put our work in context, the paper first presents other available corpora related to the speech of people with DS in section 2, in order to highlight the importance of this resource. Section 3.1 describes the video game we used for compiling the corpus. The production activities that focus on prosody are detailed in section 3.2. Section 3.3 sets out the informant profiles and section 3.4 describes the recording campaigns. Sections 4 and 5 are devoted to presenting concerns related to the evaluation of oral turns, both with manual and automatic methods. In section 6, we discuss the potential use of PRAUTOCAL for analyzing the differences between TD speakers and DS speakers, human based scores and the impact on the quality of the production mode. Section 7 refers to distribution concerns. We end the paper with the conclusions in section 8 to highlight the importance of a resource that includes different evaluations of the oral productions of speakers with DS, taking into account different aspects related to prosody, thus allowing a comparison with productions of TD speakers.

2 State of the art

There are studies about the impairments in the spoken language of people with DS that include the recording of a corpus to experimentally study various aspects of the linguistic domains, although these corpora are not publicly available (for a review see Kent and Vorperian, 2013). The aims, theoretical background and methodology of each study determine the features of the recorded corpus: content, number of speakers, demographics of the speakers, etc. Broadly speaking, there are works that focused on acoustic analysis (Albertini et al., 2010; Rochet-Capellan and Dohen, 2015; Bunton and Leddy, 2011; Seifpanahi et al., 2011; O'Leary et al., 2020), voice quality (Rodger, 2009; Lee et al., 2009), prosodic skills (Zampini et al., 2016), speech disfluencies (Eggers and Van Eerdenbrugh, 2017), speech intelligibility (Wild et al., 2018), consonant, vowel and word duration (Brown-Sweeney and Smith, 1997) and perceptual evaluations (Moura et al., 2008; O'Leary et al., 2020). There are also corpora built specifically for a task: the Alborada-I3A corpus (Saz et al., 2010), designed for research in speech technologies, and the AD-Child.Ru corpus (Lyakso et al., 2019), created for studies of speech development.

A summary of the corpus used by studies about DS speech is shown in Table 1. With regard to languages, most of the works focused on the English language (Rodger, 2009; Lee et al., 2009; Bunton and Leddy, 2011; Brown-

Sweeney and Smith, 1997; Wild et al., 2018; O'Leary et al., 2020). However, there are also some studies in Italian (Zampini et al., 2016; Albertini et al., 2010), Spanish (Saz et al., 2010), French (Rochet-Capellan and Dohen, 2015), Dutch (Eggers and Van Eerdenbrugh, 2017), Portuguese (Moura et al., 2008), Russian (Lyakso et al., 2019) and Farsi (Seifpanahi et al., 2011). The age of the speakers also varies, and, as a consequence, needs to be taken into account: the studies recorded only adults (Lee et al., 2009; Rochet-Capellan and Dohen, 2015; Bunton and Leddy, 2011; Seifpanahi et al., 2011), only children (Zampini et al., 2016; Eggers and Van Eerdenbrugh, 2017; Moura et al., 2008; Brown-Sweeney and Smith, 1997; Lyakso et al., 2019) or both adults and children (Albertini et al., 2010; Rodger, 2009; Saz et al., 2010; Wild et al., 2018; O'Leary et al., 2020). Like ours, most of the recorded corpora included a control group of TD people in order to compare the spoken material with that of the DS group. Concerning content, different speech categories were recorded: vowels (Seifpanahi et al., 2011); words (Albertini et al., 2010; Brown-Sweeney and Smith, 1997; Wild et al., 2018); words and sentences (Saz et al., 2010; Bunton and Leddy, 2011); vowels, read and spontaneous speech (Lee et al., 2009; O'Leary et al., 2020); picture-description exercises (Rodger, 2009); semi-structured play sessions in interaction with participants' mothers (Zampini et al., 2016); vowel-consonant-vowel bisyllables (Rochet-Capellan and Dohen, 2015); play sessions with a toy or book (Eggers and Van Eerdenbrugh, 2017); words, spontaneous speech, read speech, play with toys, picture description, story retelling (Lyakso et al., 2019); vowels and naming figures presented on cards (Moura et al., 2008).

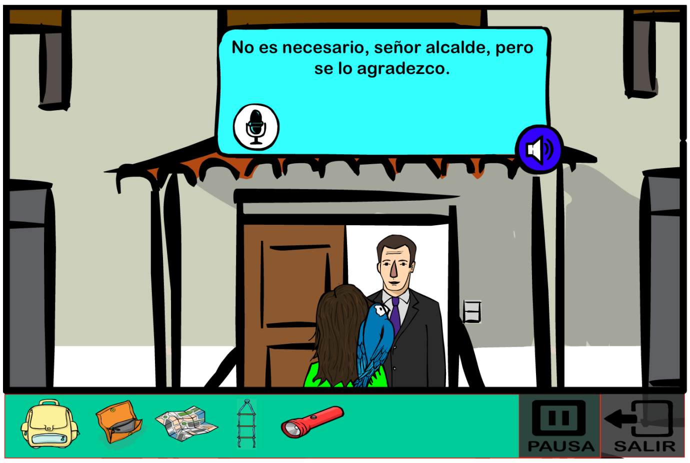
Other disorders and pathologies that also affect spoken language and communication skills, such as dysarthria, aphasia, autism, Parkinson's disease, dementia, Alzheimer's disease, etc, have also been investigated and spoken corpora have been collected. The reference corpora for dysarthric speech in American English include Nemours (Menendez-Pidal et al., 1996), Universal Access (Kim et al., 2008) and TORGO (Rudzicz et al., 2012). There are also corpora in other languages, such as Korean (Kim et al., 2016) and French (Fougeron et al., 2010; Meunier et al., 2016). From an applied approach, projects on speech technology recorded a corpus of dysarthric speakers operating their home appliances using voice commands (Nicolao et al., 2016; Gemmeke et al., 2013). AphasiaBank (Forbes et al., 2012; MacWhinney et al., 2011) is a reference corpus for the study of aphasia. Other corpora of aphasic speech are designed for automatic speech intelligibility assessment (Le et al., 2016), and contain interactions with a tablet application designed for therapeutic purposes. As far as the autism spectrum disorder is concerned, speech corpora based on recording the sessions of the Autism Diagnostic Observation Schedule (ADOS) have been collected (Lahiri et al., 2020; Li et al., 2019; Lin et al., 2018), an instrument for the diagnosis and assessment of autism. There are also corpora recorded with the aim of identifying and assessing the severity of the speech disorder in patients with Parkinson's disease (Hauptman et al., 2019; Khan et al., 2020). Several databases for studying the speech of patients with dementia are available. DementiaBank contains the recordings

	Type of	#Speakers		Recorded	Units per	
Authors	**speakers**	**DS**	**TD**	**units**	**speaker**	**Language**
Brown-Sweeney and Smith (1997)	Children	16	16	Words	12 words x 7 times	English
Moura et al. (2008)	Children	66	204	Vowels Semi-spontaneous	5 vowels x 5 times several figure names	Portuguese
Rodger (2009)	Adults Children	22	52	Semi-spontaneous	5 picture descriptions	English
Lee et al. (2009)	Adults	9	9	Vowels, Reading, Natural speech	3 vowels, 1 reading, 1 minute	English
Saz et al. (2010)	Adults Children	3	232	Words Sentences	(57 words 28 sentences) x 4 times	Spanish
Albertini et al. (2010)	Adults Children	78	106	Words	NA	Italian
Bunton and Leddy (2011)	Adults	2	2	Words Sentences	53 monosyllabic words, 6 sentences	English
Seifpanahi et al. (2011)	Adults	22	22	Vowels	vowel /a/ x 5 times	Farsi
Rochet-Capellan and Dohen (2015)	Adults	8	8	Vowel-consonant-vowel	144 vowel-consonant-vowel	French
Zampini et al. (2016)	Children	9	12	Semi-spontaneous	20 minutes	Italian
Eggers and Van Eerdenbrugh (2017)	Children	26	0	Semi-spontaneous	15 minutes	Dutch
Wild et al. (2018)	Adults Children	62	25	Words	20 words	English
Lyakso et al. (2019)	Children	24	80	Read speech Semi-spontaneous Spontaneous	NA	Russian
O'Leary et al. (2020)	Adults Children	3	3	Vowels, Words, Sentences, Spontaneous speech	NA	English

Table 1 Description of other previous corpora that are focused on the study of oral productions of speakers with DS (NA: not available).

from English speakers with Alzheimer's disease during a picture description task (Becker et al., 1994) in a five year longitudinal study. ILSE is a German database that includes participants with age-associated cognitive impairment and Alzheimer's disease, recorded over the course of 20 years (Weiner et al., 2016). The Hungarian MCI-mAD database contains recordings of speakers with cognitive impairment and Alzheimer's disease (Gosztolya et al., 2019). A corpus of French patients with dementia was recorded to study the automatic assessment of dementia (Satt et al., 2014).

In this context, the use of speech enabled games can help in the recording process, motivating the participant speakers. For instance, Voice Race is an online educational game that has been used to collect an English corpus of over 55,000 utterances (18.7 hours of speech) (McGraw et al., 2009). It is based on the use of flashcards, and players earn points by using speech to match terms with their definitions. Another example is the quiz game that was used to collect 18,300 utterances (3.87 hours of speech) of European Portuguese speech over the Internet (Freitas et al., 2010). Finally, a web enabled multimodal language game has been used to collect an English-L2 speech corpus produced by Swiss German-L1 students. The recorded corpus contains 814 utterances in which the subjects had conversations with an animated agent, obtaining points and badges (Baur et al., 2014). In view of the goodness of using speech enabled



No es necesario, señor alcalde, pero se lo agradezco.

Fig. 1 Screenshoot of the video game. The scene refers to a production activity

games to collect speech samples, we used the same approach for compiling a corpus of speech in DS.

3 Data collection methods

3.1 The learning video game

As detailed in González-Ferreras et al. (2017) and Aguilar (2019), the video game has the structure of a graphic adventure game, providing conversations with characters and navigation through scenarios. The video game includes three types of activities: comprehension, production and visual cognitive activities. Comprehension activities are focused on lexical-semantic comprehension and the improvement of prosodic perception in specific contexts. In these activities, players must choose between different options to continue a conversation with a game character. Production activities are aimed at oral production, so the player is encouraged by the game to train his/her speech, keeping in mind such prosodic aspects as intonation, expression of emotions or syllabic emphasis (see section 3.2.1). In these activities, the player engages in conversation with the characters and must choose between different options to continue the dialogue or to record some sentences (see Figure 1). These activities are the source of the oral corpus that we describe in this paper. Finally, the visual activities are included to add variety to the game and to practice other skills not directly related with language, such as attention or visual perception skills.

Information about the user interaction and the audio recordings of the production activities are stored automatically by the video game during the game sessions. This information can be used to analyze the user's evolution in successive game sessions and the audio recordings increase the speech corpus. This user interaction log has information about the game time, the attempts to complete a task, the number of mouse clicks or the assistance given to the user. As a result of the evaluation of the proficiency of the user performance during the different activities, the player is allowed to continue playing or forced to repeat the current activity. A maximum number of attempts was fixed at three to avoid frustration in the players. In addition, to reduce the ambient noise in the recording process, the players used a headset with microphone incorporated (Plantronics USB headset, recording MS-WAVE PCM audio with frame rate 44100, 16 bits per sample and mono).

We used three different versions of the same video game to compile the corpus: the source version, called *La piedra mágica*, whose architecture is described in González-Ferreras et al. (2017); a more elaborated and complete release called PRADIA (Aguilar, 2019), with the same game narrative, but with an increased number of games and tasks; and a simplified version that includes only the production activities that were used to collect audio samples from TD users. The current version of PRADIA can be used in three modalities, depending on how the quality of the spoken answers is evaluated to allow the player to continue: automatic decision, three-level human based decision or template-based human based decision (detailed in section 4.1.2).

3.2 Speech production activities

The prosodic categories follow the framework of intonative phonology, which takes into account the difference between linguistic and paralinguistic categories in the prosody component. From this approach, we assume the existence of prosodic categories (prominence, organization of sentences in prosodic groups and intonation), which can be modified in their phonetic implementation depending on paralinguistic cues (e.g., emotions) (Ladd, 2008). Therefore, prosody is conceived as a phenomenon of both form and meaning, and prosodic differences, regardless of their systematic, conventional or natural nature, affect the processes of meaning interpretation. In what refers to linguistic functions, we follow Halliday's categories (Halliday, 1970), according to the particular paralinguistic situation in which the oral turn of the user is produced. As detailed in section 3.2.3, the production mode can change as the video game adapts to the speakers' difficulties when they have to repeat the activities.

Both prosodic and linguistic functions are a matter of interest in the study of DS language development. Stojanovik (2011) and Loveall et al. (2021) showed that DS speakers present deficits in prosodic production affecting focus, chunking and turn-end. Concerning linguistic functions, Abbeduto et al. (2007, 2008) report how DS speakers display areas of substantial pragmatic weakness, such as rendering descriptions and interacting. Martin et al. (2009) highlight the fact that expressive language skills are more impaired than receptive skills in young individuals with DS when referring to pragmatics.

Table 2 presents the list of production activities, labeled in terms of production mode, and the prosodic and linguistic features that can be trained. In order not to lose the playful component of the tool and its degree of playability, a small set of representative sentences of each category has been chosen, due to the limited number of activities that can be included in the video game. The table reports the different competences that are trained in each of the production activities of the learning game (comprehension and visual cognition complement the list of competences). The table is especially useful for speech therapists to select the activity according to the needs of the student. In the following subsections, we explain the meaning of the different aspects depicted in the columns.

3.2.1 Prosodic function

Prosodic function includes intonation, prominence and the occurrence of prosodic boundaries (phrasing). In the modality column, the different intonational patterns related to the grammatical structure are considered: D stands for declarative sentences, Q for questions (including wh-questions and yes-no questions), E for exclamations (including commands, offers and invitations).

The column Boundaries refers to the phrasing, that is, the organization of fluent speech into groups. Although there is no agreed account of the prosodic domains in Spanish, we distinguish two main ones: the intonation phrase and the intermediate phrase, which can be differentiated by their degree of prosodic autonomy. The arguments supporting the distinction come from perception and tonal inventory. On the one hand, speakers easily discriminate two levels of prosodic separation, normally associated with the finality or non-finality meaning of the sentence (Estebas-Vilaplana et al., 2015); on the other hand, the intermediate phrase is tonally marked with an accent, but the inventory of boundary tones that can appear in this position is more restricted than that which marks the end of an intonation phrase (Aguilar et al., 2009).

Prominence in speech can be conceived from several dimensions: phonetics, phonology, semantics, pragmatics, etc. (Cole, 2015). In this study, we adopt the perspective according to which prominence expresses the meaning of a statement in relation to the discursive context; in particular, how it serves to point out relevant information in statements, highlight an element or correct some information. The Prominence column classifies the target sentences of the activities according to word (W) or sentence (S) prominence. A word is prominent when it has been pronounced with prosodic salient features that put it in a noticeable position in the sentence. The whole sentence exhibits prominence (S) in those cases with a rhetorical and grandiloquent speech production within the video game context.

Although each target sentence is representative of one of the categories in the table, each of them also serves to observe the player's phonetic and phonological performance with respect to other phenomena, such as the position of the lexical accent and its phonetic realization, or the duration of syllables and pauses.

		Prosodic Function			Language Function			Prod. Mode	
Activity code	Expected utterance	Modality	Boundaries	Prominence	Function	Politeness	Emotion	First attempt	Other attempt
D0120	¡Hasta luego, tío Pau! (*See you later, Uncle Pau!*)	E	IF		S	F		R	R
D0211	¿Dónde hay libros de historia, por favor? (*Where are there history books, please*)	QD	IF		HS	C		E	I
D0220	¡Muchas gracias Juan! (*Thank you very much Juan*)	D	IF		S	T		E	I
D0310	Hola, ¿tienen lupas? Quería comprar una. (*Hello, do you have magnifying glasses. I wanted to buy one.*)	DQD	FFF		SHR	G		E	I
D0320	Sí, la necesito. ¿Cuánto vale? (*Yes, I need it. How much does it cost?*)	DQ	IFF		NH			R	R
D0420	Hola tío Pau. Ya vuelvo a casa. (*Hello, Uncle Pau. I'm coming home now.*)	DD	IFF		SR	G		E	I
D0430	Sí, esa es. ¡Hasta luego! (*Yes, that's it. See you later!*)	DE	IFF		RS	F		R	R
D0510	¡Hola, tío Pau! ¿Sabes dónde vive la señora Luna? (*Hello Uncle Pau¡Do you know where Mrs. Luna lives?*)	EQ	IFF		SH	G		R	R
D0820	Si no tardo, sí. Por favor. (*If it doesn't take long, YES please.*)	DD	IFF	W	OS	C		R	R
D0910	Hola. Necesito una escalera. (*Hi, I need a ladder.*)	D	IF		SN	G		E	I
D0930	No, no. La de CUERDA, por favor. (*no, no, the ROPE one please*)	D	IIF	W	OS	C		E	I
D0940	No, eso es todo. Gracias. (*No, that's all. Thank's*)	DD	IFF		RS	T		R	R
D1110	Buenos días, Señora Molina. ¿Está Pedro en casa? . (*Good morning, Miss Molina. Is Pedro at home?*)	DQ	IFF		SH	G		R	R
D1120	Buenos días, Señora Molina. (*Good morning, Miss Molina.*)	D	IF		S	G		I	I
D1130	¿Está Pedro en casa? (*Is Pedro at home?*)	Q	F		H			I	I
D1140	De acuerdo. Muchas gracias. (*All right. Thank you very much.*)	DD	FF		RS	T		R	R
D1210	¡Ey, Pedro! ¿Cómo estás? (*Hey Pedro, how are you?*)	EQ	FF		SH	G		R	R
D1220	Ojalá pudieras ¿Me dejas tu linterna? (*I wish you could, can I have your flashlight?*)	DQ	FF		PH		D	E	I
D1230	¿Me dejas tu linterna? (*Can I have your flashlight?*)	Q	F		H			E	
D1240	Me tengo que ir ya, Pedro (*I have to go now, Pedro.*)	D	IF		R			E	I
D1510	Soy quien busca la piedra mágica. Necesito ver al alcalde. (*I am the one who seeks the magic stone. I need to see the Mayor.*)	DD	FF		RN			R	R
D1511	Soy quien busca la piedra mágica. (*I am the one who seeks the magic stone.*)	D	F		R			R	R
D1512	Necesito ver al alcalde. (*I need to see the mayor*)	D	F		N			R	R
D1520	¿Sabe cómo ir a su casa, señor? (*Do you know how to get his house, sir?*)	Q	IF		H			R	R
D1610	No es necesario, señor alcalde, pero se lo agradezco. (*It is not necessary, Mr. Mayor, but I thank you.*)	D	IIF		S	T		R	R
D1710	Es que NO sé dónde está la piedra mágica. (*I just DON'T know where the magic stone is.*)	D	F	W	R			R	R
D1720	Tranquilo. Yo le ayudaré con mucho gusto. *It's okay. I will be happy to help you.*)	DD	FF		OR	R		R	R
D1730	Sí, claro. Aquí está. (*Yes, of course. Here it is.*)	DD	IFF		SR	C		R	R
D1740	Muchas gracias, señora alcaldesa. (*Thank you very much, Madam Mayor.*)	D	IF		S	T		R	R
D1900	Sí. El alcalde me ha dado esta tarjeta. (*Yes. The Mayor gave me this card.*)	DD	FF		RR			R	R
D2000	Seguro que sí. Buenas noches, Lolo. (*I'm sure you do. Good night, Lolo*)	DD	FIF		IS	F		R	R
D2200	Buenos días. ¡Qué triste estar solo en este bosque! (*Good Morning. How sad to be alone in this forest!*)	DE	FF		SP	G	S	R	R
D2400	SOlo la PIEdra de FUEgo Abre la PUERta del TEMplo. (*Only the fire stone opens the temple door.*)	D	F	S	I			R	R
D2500	¡Hala! ¡Esa puerta no estaba aquí antes! *Wow! That door wasn't here before!*	EE	FF		PR		O	R	R
D2600	No sé qué piedra elegir... (*I don't know which stone to choose ...*)	D	F		P	D		R	R
D2900	Ábrete PUERta y que BRIlle la PIEdra. (*Open up, door, and let the stone shine.*)	E	IF	S	O			R	R
D3000	free speech				RP		H	S	
D3100	Muchas gracias. Lo recordaré (*Thank you very much, I will remember it.*)	DD	FF		SR	T		E	I
D3400	Hola, ¿cómo estás?. (*Hi how are you doing?*)	DQ	IF		SS	G		E	I
D3700	¿Me da un billete, por favor?. (*Can I get a ticket, please?*)	QD	IF		OS	C		E	I

Table 2 Prosodic and pragmatic categories trained in the production activities of the video game and strategies to obtain the target sentences. *Modality* distinguishes declarative (D), questions (Q) and exclamatory sentences(E); *Boundaries* intermediate (I) and final (F) boundaries; *Prominence* highlights words (W) or sentences (S) to be pronounced in a salient way; *Function* distinguishes instrumental (N), regulatory (O), interactional (S), personal (P), heuristic (H), imaginative (I) and representational (R) functions; *Politeness* distinguishes greetings (G), farewells (F), thank you (T) and courtesy (C) formulas; *Emotions*, sadness (S), surprise (O), disgust (D) and happiness (H). The *Production Mode* columns distinguish imitation (I), reading (R), elicited (E) and spontaneous (S) speech.

3.2.2 Functions of language

To accommodate the learning objectives of the video game to Halliday's Theory of Language Development (Halliday, 1970), of general use in the speech and language therapy domain, we have classified the target sentences according to those functions that help the speaker to satisfy physical, emotional and social needs. Halliday refers to them instrumental, regulatory, interactional, personal, heuristic, imaginative and representational functions. Instrumental

(N in Table 2 column *Function*) is when the speaker uses language to express her/his needs; regulatory (O) when language is used to tell others what to do; interactional (S) for making contact with others and form relationships; personal (P) when language is used to express feelings, opinions, and individual identity; heuristic (H) is used to seek information and ask questions; imaginative (I) to express creativity of poetic language; and representational (R) to give information and reporting facts.

We especially focus on interactional and personal functions, in particular, verbal politeness and the prosodic expression of emotions. Verbal politeness can be considered from the perspectives of linguistic features, participants' socio-cultural background and their membership within a speech community, or it can be viewed from the ways to which it is applied in interpersonal utterances (Leech, 2016). We focus on the politeness conveyed conveyed by words. The video game includes a set of target sentences that support politeness, including speech acts (basic units of linguistic communication with which an action is performed), such as greetings (G in Table 2), farewells (F), thank you (T) and courtesy (C) formulas (Haverkate, 1988; Vidal, 1995; Van Olmen, 2017). Concerning emotions, the video game presents activities related to surprise (O in Table 2 column *Emotion*), happiness (H), disgust (D), sadness (S). Fear and anger complete the list of basic emotions (Saarni et al., 2007). They have not been included because they did not fit with the story line of the video game.

3.2.3 Production mode

The prosodic categories and linguistic functions are addressed by different procedures, depending on whether the player is required to read (R in Table 2 column *ProdMode*), imitate a previous model (I) or produce an elicitated sentence on his/her own (E). Spontaneous speech (S) is only recorded at the end of the game, when the player has solved the full adventure and is asked about his/her feelings. The differences in the tasks enrich the information provided by the use of the video game concerning the prosodic and pragmatic skills of people with DS, as the relations between oral language skills and reading skills is a main topic of research in the literature of reading difficulties (Brooks, 2013; Roch et al., 2015, 2012).

3.3 Informant profiles

Table 3 and Fig. 2 describe the characteristics of PRAUTOCAL informants. The corpus is balanced in terms of gender (49 males, 41 females) and profile (50 individuals with intellectual disability and 40 TD speakers). The age range of both types of speaker is also similar, going from 13 to 42 years in the case of speakers with DS and from 6 to 68 years in the case of TD speakers; mean ages are 21.55 and 29.72, respectively. The oral productions of a group of TD children are included in PRAUTOCAL in order to have a control group that can match the mental age of DS speakers.

Some of the participants with DS were given tests to account for individual variability by getting measurements of different developmental variables. Specifically, 5 participants of the C2 campaign and almost all of the participants (17) of the C3 campaign were evaluated using the following tests (campaigns are detailed in the following subsection). The Peabody Picture Vocabulary Scale-III (Dunn et al., 2006) was used to assess verbal mental age, the forward digit-span subtest included in the Wechsler Intelligence Scale for Children-IV (Corral et al., 2005) was used to evaluate verbal short-term memory and Raven's Coloured Progressive Matrices (Raven et al., 1993) served as a means to measure non-verbal cognitive level. The descriptive characteristics and scores obtained are shown in Fig. 2. The full PEPS-C battery in its Spanish version (Martínez-Castilla and Peppé, 2008) was also administered to participants in order to have specific measurements of prosody level. The mean percentage of success in the perception (MPercT) and in the production PEPS-C tasks (MProd) is also presented in Fig. 2.

		Gender		CA				
Type	#	M	F	min	mean	max	#recordings	campaigns
TD	30	16	14	22	37.96	68	1589	C1 C6
DS	42	24	18	13	21.55	42	2151	C1-5
OS	8	3	5	17	19.12	21	281	C1 C4 C5
TC	10	6	4	6	8.3	11	154	C1 C6

Table 3 Characteristics of informants: CA is chronological age expressed in years. DS is Down syndrome speaker, TD is typically developing adults, OS refers to individuals with another type of intellectual disability and TC are typically developing children.

The variety of intellectual profiles of the speakers with DS is high as far as verbal age (from 4.17 to 9.33 years) and non-verbal cognitive level (from 10 to 27 raw score) are concerned, as is usual in the case of DS (Chapman and Hesketh, 2001). Regarding the 5 speakers evaluated in campaign C2, the short-term verbal memory varies from 6.17 to 11.17 (mental age in years). As for prosodic competence, measured with the PEPS-C test, the corpus has individuals with diverse capabilities ranging from 50 to 84.4 (percentage of success) in the case of perception of prosody and from 31.3 to 82.8 in the case of production.

The Pearson correlation between these indices is low, the highest one being the correlation between MPrecT and MProdT (0.77). The rest of the indices presented in Fig. 2 correlate with each other, with a Pearson correlation factor below 0.51. The indices of the PEPS-C test correlate the most because they refer to the same competence, namely, prosody; the other indices correlate little, thus reflecting the high diversity of speakers with DS concerning both intellectual capabilities and language competences.

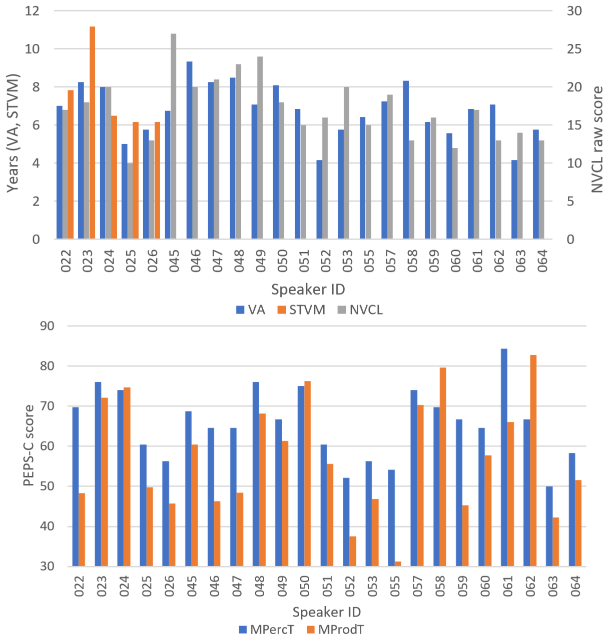


Fig. 2 Characteristics of speakers with DS: VA is verbal mental age in years (min 4.17; mean 6.8; max 9.33), STVM is short-term verbal memory in years (min 6.17; mean 7.57; max 11.17), NVCL is non-verbal cognitive level raw score (min 10.0; mean 17.13; max 27.0), MPercT is the mean percentage of success in perception in PEPS-C (min 50.0; mean 65.64; max 84.4) and MProdT is the mean production of success in production in PEPS-C (min 31.3; mean 57.33; max 82.8). Section 3.3 details the metrics and tests used to compute them. STVM was only measured in one of the collaborating centers.

3.4 Recording campaigns

PRAUTOCAL was gathered in six recording campaigns. The details of each campaign are shown in Table 4. Some users participated in multiple recording campaigns, and others only in one of them. The campaigns correspond to the evaluations of different versions of the video game which could imply different evaluation methodologies. Some of the campaigns are justified with the need of having data of TD speakers. The particularities of each campaigns are detailed in the following paragraphs.

Campaign C1, the first recording campaign, was carried out using the initial version of the video game. With the aim of detecting possible deficiencies in the user interface and to determine how the users interacted with the video game, a usability test was carried out in the first game session (Corrales-Astorgano et al., 2016; González-Ferreras et al., 2017). The rest of the sessions were only dedicated to playing the game.

During the four game sessions, the role of the trainer (a teacher or speech therapist), who sat next to the player, was twofold: on the one hand, he/she evaluated the player's recordings and on the other, he/she helped players if

Campaign	#Speakers	#Audios	Length (seconds)	TD	DS	OS	TC
C1	51	677	2070.69	21	13	7	10
C2	10	773	2705.96	0	10	0	0
C3	19	739	3045.7	0	19	0	0
C4	11	439	1600.2	0	9	2	0
C5	11	111	352.3	0	9	2	0
C6	16	1436	3858.45	15	0	0	1
Total	90	4175	13633.3	30	42	8	10

Table 4 Number of speakers, number of audios, total audio length (in seconds) and number of speakers of each type, for each recording campaign. TD means typically developed adults, DS means Down syndrome, OS means other intellectual disabilities and TC means children with typical development. There are speakers that participated in multiple campaigns, so Total means the number of different speakers

necessary. The recordings gathered in this campaign were evaluated afterwards by a prosody expert, following the criteria described in section 4.1.2.

A total of 51 users participated in the game sessions: 21 TD adults (12 males and 9 females), 13 people with DS (9 males and 4 females), 7 people with an intellectual disability of unknown origin (3 males and 4 females) and 10 TD children (6 males and 4 females). The first version of the video game included 7 production activities aimed at obtaining the sentences in Table 2.

Ten adults with DS (6 males and 4 females) participated in campaign C2. For sample selection, teachers working at the centers were asked to choose individuals with DS of different developmental levels. Participants played with the latest version of the video game (Aguilar, 2019).

As happened in the first campaign, the participants were supported by a speech and language therapist who was an expert at working with individuals with DS. The therapist explained the game, helped participants when needed and took notes about how each session developed. In addition, the therapist also assessed participants' speech production and thus monitored their rhythm of progress within the video game. The criteria followed by the therapist to evaluate the recordings are described in section 4.1.2. Not all the evaluations were overseen by a speech therapist, so these evaluations were not taken into account when building the evaluation data.

Nineteen adults with DS participated in campaign C3. The session was also carried out using the PRADIA video game but, in this case, the therapist applied a rubric (see section 4.1.3) to rate the quality of the audio recording, using a specifically designed mobile application.

Campaign C4 was carried out with the partipation of 11 people: 9 with DS (5 males and 4 females) and two females with an intellectual disability of unknown origin, but who were classmates of the DS students and shared their same learning activities. The PRADIA video game was used to obtain the recordings but, in this session, the evaluation of the recordings was made by an automatic module integrated in the video game, sending the recordings to be evaluated by a software application hosted in an external computer. The 11 users played the video game in the same room and were assisted by the

Evaluation	R	W	P	Total	#Rt.	Campaigns	Speakers
Therapist 3-level evaluation	293(14%)	155(7%)	157(7%)	605	1	C2	5
Prosody expert binary evaluation	600(28%)	366(17%)		966	1	C1-C2	23
Therapist template-based evaluation	546(25%)	193(9%)		739	2	C3	19
Binary automatic evaluation	1494(69%)	657(31%)		2151	1	C1-C5	42

Table 5 Different types of human and automatic evaluations collected in the corpus. R (Right) means move to the next activity with right as result; P (Poor) means move to the next activity but the oral activity could be better and the video game advances to the next activity; W (Wrong) means that the game offers a new attempt in which the player has to repeat the activity. The number in brackets represents the number of evaluations divided by the total number of recordings of people with DS, in percentages. #Rt is the number of raters.

therapist if help was needed. The automatic evaluation module is described in section 5.

The same speakers included in the C4 campaign participated in campaign C5. In this session, the speakers played the video game alone. During the game session, the speakers were observed by the research team and the therapist. The therapist only assisted them if they got stuck or any technical issue happened. The same automatic evaluation module of the previous campaign was used to evaluate the recordings.

Campaign C6 was done with the aim of balancing the corpus with recordings of people without any intellectual disabilities. Fifteen TD adults (8 males and 7 females) and one TD child (female) recorded all the sentences included in the video game, but using a simplified version that only showed the production activities, removing all the game elements that were not necessary to record the sentences. These changes allowed important time savings for the speakers in the recording session.

4 Manual annotations

4.1 Human judgments of quality

The final aim of our project was to incorporate an intelligent control module which allows autonomous gaming. To introduce a module that substitutes the trainer on deciding whether the user must repeat the production activities, we managed to train a supervised classifier from human based evaluations of audios collected along the different testing sessions of the tools. In the first versions of the video game, affecting campaign C2, the evaluation of the audio utterances followed a three-level criterion described in subsection 4.1.1. The limitations of these judgments for training the automatic component led us to carry out the off-line binary evaluation presented in 4.1.2. This second evaluation, affecting campaigns C1 and C2, allowed us to develop an automatic classifier to decide on the quality of the utterances. Seeking to improve the output of the classifier by providing information about the specific problems of the utterances, a new template-based evaluation was finally designed and applied,

Binary variables	Y	N
Intelligibility	707(33%)	10(0.01%)
Adaptation	699(33%)	19(1%)
Continue	546(25%)	193(9%)

Likert variables	1	2	3	4
Word omisions	7(0.01%)	23(1%)	99(5%)	573(27%)
Fluency	6(0.01%)	68(3%)	315(15%)	311(14%)
Speech rate	5(0.01%)	58(3%)	272(13%)	365(17%)
Melodic curve	5(0.01%)	61(3%)	200(9%)	429(20%)

Table 6 Number of judgments per dimension in the template-based evaluation.

as described in section 4.1.3. Also, an automatic binary classifier developed as described in section 5, was applied to the samples of every campaign.

Table 5 shows the different types of human and automatic-based evaluations of the recordings of the people with DS. These evaluations include 2151 automatic evaluations (all the recordings of the 42 DS speakers), 966 evaluations by a prosody expert (23 DS speakers from campaigns C1 and C2), 605 3-level based evaluations by a therapist (5 DS speakers from campaign C2) and 739 template-based evaluations by two therapists (19 DS speakers from C3 campaign). A detailed table including all available evaluations per audio sample is included and described in the corpus distribution.

4.1.1 Therapist 3-level evaluations

As we have already stated in Corrales-Astorgano et al. (2019), in the first version of the video game, the sessions must be carried out with the support of an external assistant who guides the player with disability in the development of the activities, if he/she needs it, and who decides if the responses are adequate enough. The evaluation was done with an external keyboard using three different options: R right (Continue with right result), P poor (Continue, but the oral activity could be better), W wrong (the game offers a new attempt in which the player must repeat the activity).

The criteria was based on the categories of intonational phonology (that is, intonation, accent and prosodic organization) (Ladd, 2008). In the production of the sentences, the following were expected: an adjustment to the target modality (declarative, interrogative, exclamatory), a difference between lexical stress (stressed vs unstressed syllables) / accent (accented vs unaccented syllables), and a plausible division in prosodic groups with an appropriate allocation of pauses.

In his/her final decision, the therapist also took into consideration the motivational and emotional status of each participant in each session. Participants with DS needed different reinforcements according to their abilities and, therefore, the therapist can make the evaluation criteria more flexible. For example, if the participant was getting bored, anxious, or frustrated, the therapist could use the category P to allow the speaker to continue with the video game to reduce any negative valence of the therapy context.

4.1.2 Prosody expert binary evaluation

To complement the therapist evaluations, an expert in prosody evaluated the recordings of 23 speakers with DS in an offline mode (C1 and C2 campaigns), following the criteria detailed in section 4.1.1. In this evaluation, the environmental conditions implied in the use of the video game by DS speakers (lack of interest or level of frustration) were not considered, and the judgments were reduced to a binary decision (Right or Wrong production). The evaluation relied on perceptive criteria, without any acoustic analysis of the sentences, and the possible deficits in the segmental component (pronunciation of sounds) were not considered. Audios were shown in sequence and the expert could listen to each utterance as many times as necessary before providing the judgment. Even in the case of speakers with serious problems of intelligibility, the main criterion was whether the intonation had been produced in a way close enough to the expected one. More details can be found in Corrales-Astorgano et al. (2019).

4.1.3 Therapist template-based evaluation

In this case, during the playing sessions, the therapist was supplied with a specific purpose mobile application, not only for deciding whether the speaker continued or not, but also for rating the quality of the audio according to a set of criteria. The figures in Table 5 corresponding to R and W are derived from the criteria Continue (Y/N).

The following criteria were established:

Intelligibility (Y/N): to indicate whether the message was intelligible or not.

Adequacy (Y/N): if the message was intelligible, the rater indicated whether the utterance corresponded with a suitable message according to the context.

Word omissions (Likert 1-4 scale): to indicate whether the speaker omitted words in the message. The scale was: (1) The user omits or changes 2 or more content words; (2) The production omits or changes 1 or more content words; (3) The user omits or changes 2 or more function words, but all the content words are sayd; (4) The production omits or changes 0 or 1 function words but says all the content words.

Fluency (Likert 1-4 scale): to evaluate the control of the interruption points by the speaker, with the following values: (1) There are inappropriate interruption points in almost all words; (2) Inappropriate interruption points in half or more words; (3) Inappropriate interruption points in a few words; (4) No inappropriate interruption points in any words.

Speech rate (Likert 1-4 scale): to penalize the anomalous slow utterances with the following scores: (1) Speech rate was too slow, (4) Speech with adequate speech rate, (2) and (3) are intermediate rates.

Melodic curve (Likert 1-4 scale): to rate how good the F0 contour is regarding the prosodic functions trained. The values for this parameter are: (1) Flat melodic curve or inadequate use of it during the entire production; (2) The user produces a small part of the melodic curve correctly; (3) The user produces almost all of the melodic curve correctly; (4) The user produces all of the melodic curve correctly.

Continue (Y/N): this is used by the therapist to let the user continue with the following game activity or to repeat the present one. The therapist could avoid using the value N depending on the state of the player: frustration, fatigue, etc.

4.2 Transcriptions

Every utterance of the DS recordings includes the literal transcription of the contents and a reference to the correct target production. The transcription format includes annotations of the disfluencies. We followed the criteria established in Shriberg (1994) to annotate disfluencies, which consider both interruption points and editing terms. Editing terms are used by speakers to correct the message and are a fundamental part of false starts and repetitions.

Both interruption points caused by a disfluency and correctly placed pauses are marked with the symbol ",". Thus the symbol "," should be interpreted as an internal pause in the utterance. Researchers can distinguish between disfluent interruption points and pauses introduced to separate intonation groups using the target reference sentence.

The editing terms of the disfluencies have been marked in between the symbols < and >. The criteria for marking the boundaries of the editing terms is that, when removed, the resulting text should be as close as possible to the target sentence. In that sense, the target sentence would play the role of the underlying fluent sentence used in Adell et al. (2012) for modeling disfluencies.

The transcriptions have been included to represent the phonetic content of the utterance. Thus, the transcriptions include many out of vocabulary words, and some of them are unintelligible. The corpus includes marks of fillers, breathing and noises that are marked with the symbol #.

To illustrate the annotation results, Table 7 shows the transcription of the activity D1900. The introduction of interruption points in diverse places is the main reason for the presence of different versions. Repetitions and false starts are very frequent: 15 out of the 22 transcriptions include editing terms (in between < and >). Fillers appear 7 times (marked with the symbol #). Section 6.4 presents an analysis of mistakes and divergences with respect to the target sentence to show the impact of the production modality on fluency.

5 Automatic annotation using machine learning

In Corrales-Astorgano et al. (2019), we presented a cross validation process to build an automatic labeling system of the quality of the utterances. The

Transcription	#	Transcription	#
sí, el alcalde, me ha dado, esta tarjeta	3	sí, el alcalde me ha dado esta tarjeta	2
sí tengo una tarjeta, que me ha dado el alcalde	1	sí, esta, la, tarjeta del alcalde	1
sí el alcalde me ha dado esta tarjeta	1	eh sí, el alcalde me ha dado esta tarjeta	1
sí, <el>el alcalde me ha dado esta tarjeta	1	sí, al,calde me ha dado, esta, tarjeta	1
sí, el alcalde, me ha dado, esta, tarjeta	1	sí, el, al,calde me, ha, dado, esta, tarjeta	1
sí, el, alcalde, me ha, dado, esta tarjeta	1	sí, el alcalde me ha, dado esta tarjeta	1
sí, el alcalde me ha dado <,e,> esta tarjeta	1	sí, alcalde me ha, esta tarjeta	1
sí, alcalde me ha dado esta tarjeta	1	el alcalde <me> me ha dado	1
esta es la tarjeta del, <a>alcalde	1	y tengo, esta tarjeta, que me ha dado, el alcalde	1
sí, el alcalde me ha dado, esta tarjeta	1	sí, esta tarjeta del alcalde	1
sí, tengo <est> esta tarjeta que me ha dado el alcalde	1	sí, tengo <e>tengo esta, tarjeta que ma que me ha dado el alcalde	1
sí le quiero na tane	1	sí le que, tera la alcalde	1
sí, es que ta el alcalde	1	sí, le quiero el alcalde	1
<#eh>sí, el tarjeta del alcalde	1	dame esa tarjeta, por favor	1
sí, muchas gracias	1	muchas gracias, ya tengo el, carnet	1
<#ehh> sí el, alcalde me ha dado, esta tarjeta	1	sí el, <al>alcalde me ha dado esta, tarjeta	1
<s>sí, el alcalde me ha dado esta tarjeta, tarjeta sí	1	sí, el alcalde, me ha, dado, esta tarjeta	1
el alcalde, me ha, dado, la tarjeta	1	sí <,#ehh prefiero así, prefiero así, crece sí, crece sí> alcalde, me ha dado, esta tarjeta	1
<sí, el alcalde me ha dicho,> sí, el alcalde, me ha dado, esa tarjeta	1	sí, el alcalde <,ma> me ha dado, esta tarjeta	1
sí, el alcalde, me, ha esta, tarjeta	1	sí, el alcalde, me ha dado <una,> esta tarjeta	1
sí el, alcalde, me ha dado, esa tarjeta	1	<sí, el alcalde, sí #e #e sí haber,> sí el alcalde, me ha dado, esta tarjeta	1
sí el, alcalde me ha dado esta tarjeta	1	sí, esta, tarjeta	1

Table 7 Different transcriptions of the productions of activity D1900 (expected target sentence: "Sí. El alcalde me ha dado esta tarjeta" *Yes. The Mayor gave me this card.*). # is the number or occurrences of the same transcription.

openSmile toolkit (Eyben et al., 2013) was used to extract acoustic features from each recording and the GeMAPS feature set (Eyben et al., 2016) was selected, due to the variety of acoustic and prosodic features contained in this set, which includes frequency related, energy related, spectral and temporal features. The arithmetic mean and the standard deviation normalized by the arithmetic mean were calculated on these features. Furthermore, four additional temporal features were added: the silence and sounding percentages, silences per second and the mean silences. The silence and sounding intervals were calculated using the default values of the Praat software (Boersma, 2006), excluding the initial and final silence intervals from this parameterization procedure. The complete description of these features can be found in previous research (Corrales-Astorgano et al., 2018). We also used feature selection before training the classifier: the features were selected by measuring the information gain of the training set and discarding the ones in which the information gain equalled zero.

Since the final aim of the automatic module of the video game is to decide whether the user can continue the game or should repeat the activity, we trained a binary classifier with Right (R) or Wrong (W) outputs. The Weka machine learning toolkit (Hall et al., 2009) was used to compare the performance of different types of classifiers: the C4.5 decision tree (DT), the multilayer perceptron (MLP) and the support vector machine (SVM). The stratified 10-fold cross-validation technique was used to create the training and testing datasets, as presented in Corrales-Astorgano et al. (2019). Classifiers was trained and tested with different versions of human labels, concluding that the best classification results are obtained with the SVM classifier (input: 21 preselected features including activity id; output: the human judgments assigned by the prosody expert described in section 4.1.2). Having selected the input features and the target output, a new SVM classifier was re-trained using all the prosody expert labels and included in the current version of the video game. The automatic labels generated with this classifier have been included in the corpus, resulting in 69% as R and 31% as W, as reported in table 5.

6 Discussion and potential uses of the corpus

The results presented in Table 3 and Fig. 2 show that the heterogeneity of the target population is reflected in the corpus, with important ranges of variation in the different characteristic values and moderate correlation between them. As already shown by Cleland et al. (2010), the language development level of people with DS cannot be explained only in terms of their cognitive level. For example, obtaining better results in the non-verbal reasoning test does not necessarily imply better results in prosodic production tests: in our corpus, correlation between non verbal cognitive level (NVCL) and prosodic production competence (MProdT) is close to zero. Furthermore, language deficits have projections in different aspects (one of which is prosodic production). Even informants with relatively good verbal skills can obtain low results in prosody; for example 061 obtains the best results in the MPercT test (84.4), but a moderate result in the VA test (6.1 in a range that goes from 4.17 to 9.3).

Our video game is a useful resource for corpus collection purposes, not only because the user's motivation is kept, but also because oral activities are controlled, so the collected audios can be classified a priori in terms of how useful they could be for the study the way DS speakers use different prosodic and language functions. The recording process permits 2151 sentences to be collected in controlled conditions, avoiding informant fatigue. Unlike the high workload reported in previous works, therapists reported enjoyment in the users. The fact that the production activities are framed in a gamified story that is meaningful for students allows them to maintain a high degree of attention.

In the following subsections we highlight three aspects in which the corpus can shed light on concerns that are still open in the field: the contrast between speech in DS and TD individuals, the categorization of the speaker with DS in terms of his/her speech proficiency, and the impact of the production mode in the quality of the oral productions. Our goal here is to show how the corpus could be useful for these concerns. We do not go into any of these topics in any detail as they could be a subject of analysis in future research studies. In the preliminary analysis that we present, we stress the value of the extra information included in the corpus that can be useful regarding the aforementioned concerns: prosodic/acoustic features of the utterances to show a clear difference between TD speakers and speakers with DS (subsection 6.1); manual annotation of the quality of the utterances to show possible types of speech proficiency in individuals with DS (subsection 6.2); there exists a consistent relation between automatic labels and subjective scores (subsection 6.3) and the transcriptions of the utterances to show the impact of the production mode in the quality of speech (subsection 6.4).

6.1 Comparing typically developing and Down syndrome speakers

In Corrales-Astorgano et al. (2018), we showed the importance of analyzing audio recordings of DS speakers by using the acoustic signal to complement studies based on perception. We highlighted the relative importance of the different types of prosodic features on the characterization of this type of speakers. The present corpus is twice the size of the corpus used in the previous study; we repeat here part of the experiments to show the consistency of the results.

Table 8 shows the classification results in the task of identifying the group of the speaker (TD or DS) of each utterance. In the TD group, only the samples of the adult speakers were selected. The acoustic features of the GeMAPS feature set were used (see section 5) to train two automatic classifiers: the Linear Discriminant Analysis (LDA) and the Support Vector Machine (SVM). The scikit-learn library of the python language (Pedregosa et al., 2011) was used to train these classifiers and the Recursive Feature Elimination (RFE) process was used to automatically select those features in the data that contribute most to the prediction of the speaker group. For each feature type, the precision, recall and F1-score were obtained using the classification results of each classifier. We applied the Leave-One-Subject-Out cross validation (LOSO-cv) (Sakar et al., 2013) technique to calculate these results. This technique consists of training the classifier using the data of all the speakers except one and testing this classifier with the data of the speaker that was left out of the training data. This process was applied for all the speakers. The results were similar in both classifiers, but LDA works better with the Frequency (0.72 F1-score) feature set, while SVM is better with the Energy (0.79 F1-score) and All (0.93 F1-score) feature sets. The results are the same in Temporal (0.75 F1-score), Spectral (0.79 F1-score) and Frequency+Energy+Temporal (0.86 F1-score) feature sets.

	Features		Precision score		Recall score		F1 score	
	Total	Selected	LDA	SVM	LDA	SVM	LDA	SVM
Frequency	12	4	0.73	0.72	0.73	0.72	0.72	0.71
Energy	12	12	0.78	0.79	0.78	0.79	0.78	0.79
Temporal	11	6	0.75	0.75	0.75	0.75	0.75	0.75
Frequency + Energy + Temporal	35	35	0.86	0.86	0.86	0.86	0.86	0.86
Spectral	57	33	0.79	0.79	0.79	0.79	0.79	0.79
All	92	60	0.92	0.93	0.92	0.93	0.92	0.93

Table 8 Classification results for identifying the group of the speaker. Precision, recall and F1-score using different feature sets and two different classifiers are reported. The classifiers are Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). *Features* is the number of input features in each set.

In addition, when all features are jointly considered, the best classification results are obtained, regardless of the classifier. When input features are used in isolation, frequency related features lead to the worst classification performance. Using frequency, energy and temporal features together gives a noticeably better performance than using any of these groups separately. Spectral features lead to a worse performance than using either the frequency+energy+temporal group or the complete set of features altogether. These results are similar to those obtained in previous works (Corrales-Astorgano et al., 2018).

These results show that the acoustic features can be used to discriminate between TD and DS speakers, since the distributions of their respective acoustic features are separable. PRAUTOCAL can be useful to analyze the specific characteristics of the speech of people with DS in greater depth, with the aim of improving the individual diagnosis of possible speech and language disorders that can be trained by speech therapy.

6.2 Heterogeneity of speakers with Down syndrome

Variability in the different linguistic skills of individuals with DS has often been documented (Fidler and Nadel, 2007). In Corrales-Astorgano et al. (2019), we showed how this variability affects the automatic prediction of the quality of the speakers' productions for those of campaign C2. In campaign C4, we introduced an evaluation template to deepen the analysis of the heterogeneity by using the prosodic dimensions presented in section 4.1.3. Here we show how these dimensions allow users to be clustered into different categories according to their speaking proficiency.

Table 9 presents the number of times (in percentage) the users who participated in campaign 4 obtained the highest score during the training session. Scores were introduced by the therapist following the procedure described in section 4.1.2.

We applied k-means to cluster the speakers into classes. The best compromise between number of classes and distance to the centroids was obtained for k=4. Table 9 shows the centroids and Fig. 3 presents the inter-participant dis-

Spk	Class	Melodic curve	Fluency	Omissions	Speech rate	Intelligibility	Adequacy
045	G2	72	46	85	49	100	95
046	G1	59	43	82	73	100	100
047	G2	57	40	81	40	100	95
048	G3	79	61	78	77	100	98
049	G3	95	60	83	60	100	98
050	G3	100	74	100	93	100	100
051	G2	68	62	78	41	100	97
052	G4	19	22	39	11	97	100
053	G1	57	37	80	60	94	100
055	G1	65	47	71	78	100	100
057	G3	71	63	95	92	100	95
058	G1	80	30	100	90	100	100
059	G1	68	47	85	67	98	96
060	G4	26	11	77	20	100	97
061	G3	71	77	100	74	100	100
062	G2	88	9	81	25	100	97
063	G4	16	12	85	9	85	85
064	G4	24	47	79	21	100	100
094	G2	62	38	90	27	100	100

Class	Melodic curve	Fluency	Omissions	Speech rate	Intelligibility	Adequacy
G1	66	41	84	74	98	99
G2	69	39	83	36	100	97
G3	83	67	91	79	100	98
G4	21	23	70	15	96	96

Centroids

Table 9 Automatic classification of speakers using k-means. The table on the right presents the centroids of the four classes G1-G4 and the table on the left the class assigned to each of the speakers (Spk). Figures in the cells are the percentages of rates with the highest score per criterion.

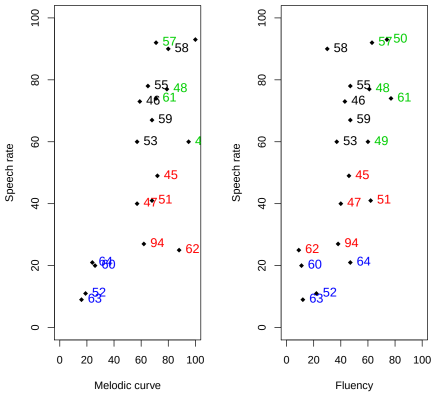


Fig. 3 Relative position of speakers using *speed* and *curve* in the axes. The clusters presented in Table 9 are coded with colors: G1 in black, G2 in red, G3 in green, G4 in blue.

tances by using the two most discriminant dimensions: *speech rate* and *melodic curve*.

The G3 group represents speakers with the highest performance in all the dimensions; at the other extreme, the G4 group of speakers obtains the worst results in all the dimensions; G2 speakers have an inter-medium level with problems concerning *speech rate*; G1 speakers are similar to the G4 speakers but they seem to be less fluent.

The distribution of participants into classes reflects the high variety of production problems that characterize speakers with DS. These preliminary results show how useful human based annotation can be to identify the specific problems of the users. Further research into this is necessary to study how this classification can improve the effectiveness of the automatic systems that could be trained to predict the quality of oral productions.

6.3 Comparing automatic and manual annotations

The automatic annotations split the corpus into two classes according to the binary R/W label assigned. Concerning the relation between automatic and manual annotations, higher human scores are in the R group. The Chi-square test of independence was used to test whether the automatic annotations are independent of the manual ones, resulting in a p-value < 0.05, except in the case of the label. We hypothesize that the combination of different human scores can shed light on the reasons for the binary category W to be assigned, but this should be analyzed further in future research.

All the samples of the corpus have automatic labels generated with the same automatic classifier that is incorporated in the current version of the video game. This fact implies the limitation that part of the samples have been used to train the automatic system. We took this decision to guarantee that all the samples have a common reference and repeatable results, but it must be taken into account by future users of the corpus. We assume that the corpus is not a closed resource and additional automatic labels will be incorporated in future versions.

Building an alternative expert system that does not limit its judgments to binary R/W decisions, but complements these judgments with information about the fluency, intonation and speed of the utterances is also left for future work. This is a challenging task because the dependence both on the prosodic and language functions and on the production mode must be taken into account. We expect that the availability of this new open language resource, in combination with other sources of information such as unsupervised labels to be included in future work, will shed light on this promising research line.

6.4 Impact of the production mode on disfluencies

The relations between oral language skills and reading skills is a major topic of research in the literature of reading difficulties (Brooks, 2013), mainly due to their implications for school intervention practices in reading training. More specifically, in individuals with DS, the study of the relations between their oral language and their reading skills is a main field of research (Roch et al., 2015). It seems that their reading comprehension is better than their oral comprehension, and their oral language skills are worse than expected on the basis of their reading comprehension skills; these differences could be due to the different allocation of verbal memory resources (Roch et al., 2012). In that sense, we believe that the data obtained in this study through reading and

Production mode	Rec	IP	ET	ETw	Fil	#I	#D	#S
Read	1463	1.62	0.37	0.48	0.10	0.38	0.83	0.56
Elicited	491	0.89	0.30	0.45	0.12	0.78	2.26	1.28
Imitation	162	1.36	0.45	0.57	0.18	0.81	1.51	0.97
Spontaneous	35	1.60	0.57	0.91	0.29	7.14	0.00	0.00
All	2151	1.43	0.36	0.48	0.12	0.62	1.19	0.74

Table 10 Mean number of lexical deviations with respect to the expected output of the production activities in the DS transcriptions. Rec is the number of recordings, IP is the mean number of interruption points per recording, ET is the mean number of editing terms per recording, ETw is the mean number of words in the editing terms per recording, Fil is the number of fillers per recording, #I, #D and #S are the mean number of insertions, deletions and substitutions per recording with respect to the expected utterance.

non reading activities would be a valuable resource to analyze the possible differences and contribute to a better understanding of the relations between these two language processes in DS.

Table 10 summarizes the disfluencies found in the recording transcriptions of the DS group in campaigns C1 to C5, showing a clear impact of the production mode on the type and quantity of the disfluencies transcribed. The production mode depends directly on the kind of activities and not on the campaign or groups of users. In descriptive terms, elicited speech shows the lowest number of interruption points (0.85 vs more than 1.44). Spontaneous speech and imitation introduce more editing terms and fillers than elicited and read speech (for example for ET 0.52 and 0.49 vs 0.38 and 0.30). Utterances produced in read mode are more similar to the expected ones than the rest of the utterances, with a lower number of insertions, deletions and substitutions (#D and #S are 0 in spontaneous speech as there is no reference).

7 Corpus description and distribution

The corpus content is organized in folders with the audio files, transcriptions, evaluations (both human and automatic), acoustic features and silent pauses of each of the utterances. The name of the files includes anonymized information of the speaker (id, gender, type and location) and information for identifying the activity, campaign and repetition. Additionally, the corpus distribution includes information about the activities, campaigns and intellectual profile of the speakers (the one reported in this paper). Finally, the log files of the game interaction are also included, as they include information that could be related with the reasoning or memory capabilities of the speakers or with the time of preparation before the oral productions.

Data collection design and trials were approved by the Ethics Committee of the University of Valladolid[1]. According to the rules of the committee, interested researchers should accept specific dissemination and use restrictions

[1] Resolution **PI 20-1639** of the Ethics Committee

identical to those established in the PRAUTOCAL project (TIN2017-88858-C2-1-R) for accessing the audio files. The rest of the corpus is available for research purposes free of charge under license CC BY-NC, accessible from the web page of the research group [2].

8 Conclusions

In this paper, we have presented a procedure for collecting an oral corpus based on the use of a video game with specific activities for training oral communication competences. The use of the video game has shown itself to be an efficient tool for collecting a corpus of the speech of individuals with DS, overcoming the difficulties in achieving this aim linked to the attention and memory problems of this type of informants.

The compiled corpus is a valuable resource for studying speech in DS that seems to surpass other corpora presented in the state of the art, not only in terms of the quantity of collected data, but also in terms of the quality, as it includes linguistic and functional information related to the specific oral turns and context; the evaluation judgments of the turns considering different dimensions; the transcriptions with annotations of the disfluencies; and the references of the speech of TD users for all the utterances.

We have shown how the recordings of speakers with DS have idiosyncratic characteristics when compared to those of TD speakers, when the high heterogeneity of the speech of DS speakers is represented in the corpus.

We expect the richness and variety of the information collected in the corpus will allow us to retrieve more knowledge about the specific characteristics of the speech of DS speakers. This could be used in future studies to enrich speech and language disorder diagnosis and improve specialized training activities.

References

Abbeduto L (2008) Pragmatic development. Down Syndrome Research and Practice

Abbeduto L, Warren SF, Conners FA (2007) Language development in down syndrome: From the prelinguistic period to the acquisition of literacy. Mental retardation and developmental disabilities research reviews 13(3):247–261

Adell J, Escudero D, Bonafonte A (2012) Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence. Speech Communication 54(3):459–476

Aguilar L (2019) Learning prosody in a video game-based learning approach. Multimodal Technologies and Interaction 3(3):51

Aguilar L, de-la Mota C, Prieto P (2009) Guía multimedia de la prosodia del español. Sitio web: http://prado uab cat/guia/es

[2] https://eca-simm.uva.es/en/prautocal-corpus/

Albertini G, Bonassi S, Dall'Armi V, Giachetti I, Giaquinto S, Mignano M (2010) Spectral analysis of the voice in Down syndrome. Research in developmental disabilities 31(5):995–1001

Baur C, Rayner E, Tsourakis N (2014) Using a serious game to collect a child learner speech corpus. In: Ninth international conference on language resources and evaluation (LREC)

Becker JT, Boiler F, Lopez OL, Saxton J, McGonigle KL (1994) The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology 51(6):585–594

Boersma P (2006) Praat: doing phonetics by computer. http://www praat org/

Brooks G (2013) The prerequisites for successful teaching and learning of literacy. European Journal of Education 48(4):557–569

Brown-Sweeney SG, Smith BL (1997) The development of speech production abilities in children with Down syndrome. Clinical Linguistics & Phonetics 11(5):345–362

Bunton K, Leddy M (2011) An evaluation of articulatory working space area in vowel production of adults with Down syndrome. Clinical linguistics & phonetics 25(4):321–334

Chapman R, Hesketh L (2001) Language, cognition, and short-term memory in individuals with Down syndrome. Down Syndrome Research and Practice 7(1):1–7

Cleland J, Wood S, Hardcastle W, Wishart J, Timmins C (2010) Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. International journal of language & communication disorders 45(1):83–95

Cole J (2015) Prosody in context: a review. Language, Cognition and Neuroscience 30(1-2):1–31

Corral S, Arribas D, Santamaría P, Sueiro M, Pereña J (2005) Escala de inteligencia de Wechsler para niños-IV. Madrid: TEA Ediciones

Corrales-Astorgano M, Escudero-Mancebo D, González-Ferreras C (2016) Acoustic analysis of anomalous use of prosodic features in a corpus of people with intellectual disability. In: Advances in Speech and Language Technologies for Iberian Languages: Third International Conference IberSPEECH, Springer, pp 151–161

Corrales-Astorgano M, Escudero-Mancebo D, González-Ferreras C (2018) Acoustic characterization and perceptual analysis of the relative importance of prosody in speech of people with down syndrome. Speech Communication 99:90–100

Corrales-Astorgano M, Martínez-Castilla P, Escudero-Mancebo D, Aguilar L, González-Ferreras C, Cardeñoso-Payo V (2019) Automatic assessment of prosodic quality in down syndrome: Analysis of the impact of speaker heterogeneity. Applied Sciences 9(7):1440

Dunn L, Dunn L, Arribas D (2006) Test de vocabulario en imágenes peabody. Madrid: TEA

Eggers K, Van Eerdenbrugh S (2017) Speech disfluencies in children with Down Syndrome. Journal of Communication Disorders

Estebas-Vilaplana E, Gutiérrez YM, Vizcaíno F, Cabrera M, de Gran Canaria P (2015) Boundary tones in spanish declaratives: Modelling sustained pitch. In: ICPhS

Eyben F, Weninger F, Gross F, Schuller B (2013) Recent developments in opensmile, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia, ACM, pp 835–838

Eyben F, Scherer KR, Schuller BW, Sundberg J, André E, Busso C, Devillers LY, Epps J, Laukka P, Narayanan SS, et al. (2016) The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7(2):190–202

Fidler DJ, Nadel L (2007) Education and children with down syndrome: Neuroscience, development, and intervention. Mental retardation and developmental disabilities research reviews 13(3):262–271

Forbes MM, Fromm D, MacWhinney B (2012) Aphasiabank: A resource for clinicians. In: Seminars in speech and language, Thieme Medical Publishers, vol 33, pp 217–222

Fougeron C, Crevier-Buchman L, Fredouille C, Ghio A, Meunier C, Chevrie-Muller C, Bonastre JF, Simon AC, de Looze C, Duez D, et al. (2010) The DesPho-APaDy Project: Developing an Acoustic-phonetic Characterization of Dysarthric Speech in French. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)

Freitas J, Calado A, Braga D, Silva P, Dias M (2010) Crowdsourcing platform for large-scale speech data collection. In: VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop

Gemmeke J, Ons B, Tessema NM, Van de Loo J, De Pauw G, Daelemans W, Huyghe J, Derboven J, Vuegen L, Van Den Broeck B, et al. (2013) Self-taught assistive vocal interfaces: An overview of the aladin project. Proceedings Interspeech 2013 pp 2038–2043

González-Ferreras C, Escudero-Mancebo D, Corrales-Astorgano M, Aguilar-Cuevas L, Flores-Lucas V (2017) Engaging adolescents with down syndrome in an educational video game. International Journal of Human–Computer Interaction 33(9):693–712

Gosztolya G, Vincze V, Tóth L, Pákáski M, Kálmán J, Hoffmann I (2019) Identifying mild cognitive impairment and mild Alzheimer's disease based on spontaneous speech using ASR and linguistic features. Computer Speech & Language 53:181–197

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1):10–18

Halliday MA (1970) Language structure and language function. New horizons in linguistics 1:140–165

Hauptman Y, Aloni-Lavi R, Lapidot I, Gurevich T, Manor Y, Naor S, Diamant N, Opher I (2019) Identifying distinctive acoustic and spectral features in

Parkinson's disease. Proc Interspeech 2019 pp 2498–2502

Kent RD, Vorperian HK (2013) Speech impairment in Down syndrome: A review. Journal of Speech, Language and Hearing Research (Online) 56(1):178

Khan T, Lundgren LE, Anderson DG, Nowak I, Dougherty M, Verikas A, Pavel M, Jimison H, Nowaczyk S, Aharonson V (2020) Assessing Parkinson's disease severity using speech analysis in non-native speakers. Computer Speech & Language 61:101047

Kim H, Hasegawa-Johnson M, Perlman A, Gunderson J, Huang TS, Watkin K, Frame S (2008) Dysarthric speech database for universal access research. In: Interspeech

Kim MJ, Wang J, Kim H (2016) Dysarthric speech recognition using kullback-leibler divergence-based hidden markov model. In: INTERSPEECH, pp 2671–2675

Ladd DR (2008) Intonational phonology. Cambridge University Press

Lahiri R, Kumar M, Bishop S, Narayanan S (2020) Learning domain invariant representations for child-adult classification from speech. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)

Le D, Licata K, Persad C, Provost EM (2016) Automatic assessment of speech intelligibility for individuals with aphasia. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(11):2187–2199

Lee MT, Thorpe J, Verhoeven J (2009) Intonation and phonation in young adults with Down syndrome. Journal of Voice 23(1):82–87

Leech GN (2016) Principles of pragmatics. Routledge

Li M, Tang D, Zeng J, Zhou T, Zhu H, Chen B, Zou X (2019) An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. Computer Speech & Language 56:80–94

Lin YS, Gau SSF, Lee CC (2018) An interlocutor-modulated attentional lstm for differentiating between subgroups of autism spectrum disorder. In: Interspeech, pp 2329–2333

Loveall SJ, Hawthorne K, Gaines M (2021) A meta-analysis of prosody in autism, williams syndrome, and down syndrome. Journal of Communication Disorders 89:106055

Lyakso E, Frolova O, Kaliyev A, Gorodnyi V, Grigorev A, Matveev Y (2019) AD-Child. Ru: Speech Corpus for Russian Children with Atypical Development. In: International Conference on Speech and Computer, Springer, pp 299–308

MacWhinney B, Fromm D, Forbes M, Holland A (2011) Aphasiabank: Methods for studying discourse. Aphasiology 25(11):1286–1307

Martin GE, Klusek J, Estigarribia B, Roberts JE (2009) Language characteristics of individuals with Down syndrome. Topics in Language Disorders 29(2):112

Martínez MH, Duran XP, Navarro JN (2011) Attention deficit disorder with or without hyperactivity or impulsivity in children with Down's syndrome. International Medical Review on Down Syndrome 15(2):18–22

Martínez-Castilla P, Peppé S (2008) Developing a test of prosodic ability for speakers of iberian spanish. Speech Communication 50(11-12):900–915

McGraw I, Gruenstein A, Sutherland A (2009) A self-labeling speech corpus: Collecting spoken words with an online educational game. In: Tenth Annual Conference of the International Speech Communication Association

Menendez-Pidal X, Polikoff JB, Peters SM, Leonzio JE, Bunnell HT (1996) The nemours database of dysarthric speech. In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, IEEE, vol 3, pp 1962–1965

Meunier C, Fougeron C, Fredouille C, Bigi B, Crevier-Buchman L, Delais-Roussarie E, Georgeton L, Ghio A, Laaridh I, Legou T, et al. (2016) The typaloc corpus: A collection of various dysarthric speech recordings in read and spontaneous styles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp 4658–4665

Moura CP, Cunha LM, Vilarinho H, Cunha MJ, Freitas D, Palha M, Pueschel SM, Pais-Clemente M (2008) Voice parameters in children with Down syndrome. Journal of Voice 22(1):34–42

Nicolao M, Christensen H, Cunningham S, Green P, Hain T (2016) A framework for collecting realistic recordings of dysarthric speech-the homeservice corpus. In: Proceedings of LREC 2016, European Language Resources Association

O'Leary D, Lee A, O'Toole C, Gibbon F (2020) Perceptual and acoustic evaluation of speech production in Down syndrome: A case series. Clinical Linguistics & Phonetics 34(1-2):72–91

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Raven J, Raven JC, et al. (1993) Test de matrices progresivas: manual/Manual for Raven's progessive matrices and vocabulary scalesTest de matrices progresivas. 159.9. 072, Paidós

Roch M, Florit E, Levorato MC (2012) The advantage of reading over listening text comprehension in down syndrome: What is the role of verbal memory? Research in Developmental Disabilities 33(3):890–899

Roch M, Florit E, Levorato C (2015) Follow-up study on reading comprehension in down's syndrome: the role of reading skills and listening comprehension. International journal of language & communication disorders pp 1–12

Rochet-Capellan A, Dohen M (2015) Acoustic characterisation of vowel production by young adults with Down syndrome. In: 18th International Congress of Phonetic Sciences (ICPhS 2015)

Rodger R (2009) Voice quality of children and young people with Down's Syndrome and its impact on listener judgement. PhD thesis, Queen Margaret University

Rudzicz F, Namasivayam AK, Wolff T (2012) The torgo database of acoustic and articulatory speech from speakers with dysarthria. Language Resources and Evaluation 46(4):523–541

Saarni C, Campos JJ, Camras LA, Witherington D (2007) Emotional development: Action, communication, and understanding. Handbook of child psychology 3

Sakar BE, Isenkul ME, Sakar CO, Sertbas A, Gurgen F, Delil S, Apaydin H, Kursun O (2013) Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. IEEE Journal of Biomedical and Health Informatics 17(4):828–834

Satt A, Hoory R, König A, Aalten P, Robert PH (2014) Speech-based automatic and robust detection of very early dementia. In: Fifteenth Annual Conference of the International Speech Communication Association

Saz O, Lleida E, Vaquero C, Rodríguez WR (2010) The alborada-i3a corpus of disordered speech. In: LREC

Seifpanahi S, Bakhtiar M, Salmalian T (2011) Objective vocal parameters in Farsi-speaking adults with Down syndrome. Folia Phoniatrica et Logopaedica 63(2):72–76

Shriberg EE (1994) Preliminaries to a theory of speech disfluencies. PhD thesis, Citeseer

Stojanovik V (2011) Prosodic deficits in children with Down syndrome. Journal of Neurolinguistics 24(2):145–155

Weiner J, Frankenberg C, Telaar D, Wendelstein B, Schröder J, Schultz T (2016) Towards automatic transcription of ilse—an interdisciplinary longitudinal study of adult development and aging. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp 718–725

Wild A, Vorperian HK, Kent RD, Bolt DM, Austin D (2018) Single-word speech intelligibility in children and adults with Down syndrome. American journal of speech-language pathology 27(1):222–236

Zampini L, Fasolo M, Spinelli M, Zanchi P, Suttora C, Salerni N (2016) Prosodic skills in children with Down syndrome and in typically developing children. International Journal of Language & Communication Disorders 51(1):74–83