An NLP pipeline as assisted transcription tool for speech therapists

LREC 2020
Language Resources and Evaluation Conference
11-16 May 2020

3rd RaPID Workshop: Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments

PROCEEDINGS

Dimitrios Kokkinakis, Kristina Lundholm Fors, Charalambos Themistocleous, Malin Antonsson, Marie Eckerström (eds.)

Proceedings of the LREC 2020 Workshop on:
Resources and Processing of Linguistic, Para-linguistic and Extra-linguistic Data from People with Various Forms of Cognitive/Psychiatric/Developmental Impairments (RaPID-3)

Edited by:
Dimitrios Kokkinakis, Kristina Lundholm Fors, Charalambos Themistocleous, Malin Antonsson, Marie Eckerström

ISBN: 979-10-95546-45-0
EAN: 9791095546450

Acknowledgments: This work has received support from the *Swedish Foundation for Humanities and Social Sciences* (RJ) through the grant agreement no: NHS14-1761:1 and the *Centre for Ageing and Health* (AgeCap, https://agecap.gu.se/).

For more information:
European Language Resources Association (ELRA)
9 Rue des Cordelières
75013 Paris
France
http://www.elra.info
Email : info@elda.org

RaPID3@LREC2020 - Preface

Welcome to the LREC2020 Workshop on "Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments" (RaPID-3).

RaPID-3 aims to be an interdisciplinary forum for researchers to share information, findings, methods, models and experience on the collection and processing of data produced by people with various forms of mental, cognitive, neuropsychiatric, or neurodegenerative impairments, such as aphasia, dementia, autism, bipolar disorder, Parkinson's disease or schizophrenia. Particularly, the workshop's focus is on creation, processing and application of data resources from individuals at various stages of these impairments and with varying degrees of severity. Creation of resources includes e.g. annotation, description, analysis and interpretation of linguistic, paralinguistc and extra-linguistic data (such as spontaneous spoken language, transcripts, eyetracking measurements, wearable and sensor data, etc). Processing is done to identify, extract, correlate, evaluate and disseminate various linguistic or multimodal phenotypes and measurements, which then can be applied to aid diagnosis, monitor the progression or predict individuals at risk.

A central aim is to facilitate the study of the relationships among various levels of linguistic, paralinguistic and extra-linguistic observations (e.g., acoustic measures; phonological, syntactic and semantic features; eye tracking measurements; sensors, signs and multimodal signals). Submission of papers are invited in all of the aforementioned areas, particularly emphasizing multidisciplinary aspects of processing such data and the interplay between clinical/nursing/medical sciences, language technology, computational linguistics, natural language processing (NLP) and computer science. The workshop will act as a stimulus for the discussion of several ongoing research questions driving current and future research by bringing together researchers from various research communities.

Topics of Interest

The topics of interest for the workshop session include but are not limited to:

- Infrastructure for the domain: building, adapting and availability of linguistic resources, data sets and tools
- Methods and protocols for data collection
- Acquisition and combination of novel data samples; including techniques for continuous streaming, monitoring and aggregation; as well as self-reported behavioral and/or physiological and activity data
- Guidelines, protocols, annotation schemas, annotation tools
- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, feature combination from different sources and modalities
- Domain adaptation of NLP/AI tools
- Acoustic/phonetic/phonologic, syntactic, semantic, pragmatic and discourse analysis of data; including modeling of perception (e.g. eye-movement measures of reading) and production processes (e.g. recording of the writing process by means of digital pens, keystroke logging etc.); use of gestures accompanying speech and non-linguistic behavior

- Use of wearable, vision, and ambient sensors or their fusion for detection of cognitive disabilities or decline
- (Novel) Modeling and deep / machine learning approaches for early diagnostics, prediction, monitoring, classification etc. of various cognitive, psychiatric and/or developmental impairments
- Evaluation of the significance of features for screening and diagnostics
- Evaluation of tools, systems, components, metrics, applications and technologies including methodologies making use of NLP; e.g. for predicting clinical scores from (linguistic) features
- Digital platforms/technologies for cognitive assessment and brain training
- Evaluation, comparison and critical assessment of resources
- Involvement of medical/clinical professionals and patients
- Ethical and legal questions in research with human data in the domain, and how they can be handled
- Deployment, assessment platforms and services as well as innovative mining approaches that can be translated to practical/clinical applications
- Experiences, lessons learned and the future of NLP/AI in the area

Submissions

Papers were invited in all of the areas outlined in the Topics of interest, particularly emphasizing multidisciplinary aspects of processing such data and the interplay between clinical/nursing/medical sciences, language technology, computational linguistics, NLP, and computer science. We welcomed also papers discussing problems derived from the design of relevant data samples and populations, but also the exploitation of results and outcomes as well as legal and ethical questions on how to deal with such data and make it available. Furthermore, the workshop solicited papers describing original research; and preferably describing substantial and completed work, but also focused on a contribution, a negative result, an interesting application nugget, a software package, a small, or work in progress. The workshop acted as a stimulus for the discussion of several ongoing research questions driving current and future research and challenges by bringing together researchers from various research communities.

We are grateful to our Program Committee members for their hard work in reading and evaluating all submissions. At the end, each submission received between 2 to 5 reviews, which helped the authors revise and improve their papers accordingly.

Unfortunately the workshop, which was originally planned to take place on the 11th of May 2020 in conjunction with the LREC 2020 conference, could not be held as a face-to-face meeting due to the ongoing Covid-19 pandemic. Nevertheless, there were 18 contributions accepted for the workshop (6 to be oral presentations and 12 to be posters). A keynote talk was invited by Dr. Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK, with the title: "Harnessing voice signals using signal processing and statistical machine learning: applications in mental health and other biomedical and life sciences applications".

Workshop website: https://spraakbanken.gu.se/en/rapid-2020.

Organizers:

Dimitrios Kokkinakis, University of Gothenburg, Sweden (*Workshop's chair*)
Kristina Lundholm Fors, University of Gothenburg, Sweden
Graeme Hirst, University of Toronto, Canada
Malin Antonsson, University of Gothenburg, Sweden
Charalambos Themistocleous, Johns Hopkins University, Baltimore, USA Marie
Eckerström, University of Gothenburg, Sweden

Program Committee (in alphabetic order):

Jan Alexandersson, DFKI GmbH, Germany
Malin Antonsson, University of Gothenburg, Sweden
Eiji Aramaki, Nara Institute of Science and Technology (NAIST), Japan
Visar Berisha, Arizona State University, USA
Ellen Breitholtz, University of Gothenburg, Sweden
Marie Eckerström, the Sahlgrenska Academy, University of Gothenburg, Sweden
Valantis Fyndanis, University of Oslo, Norway
Peter Garrard, St George's, University of London, UK
Kallirroi Georgila, University of Southern California, USA
Annette Gerstenberg, University of Potsdam, Germany
Katarina Heimann Mühlenbock, University of Gothenburg, Sweden
Graeme Hirst, University of Toronto, Canada
Christine Howes, University of Gothenburg, Sweden
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Alexandra König, Geriatric Hospital Nice and the University of Côte d'Azur, France
Nicklas Linz, DFKI GmbH, Germany
Peter Ljunglöf, University of Gothenburg, Sweden
Kristina Lundholm Fors, University of Gothenburg, Sweden
Saturnino Luz, University of Edinburgh, UK
Juan José García Meilán, Universidad de Salamanca, Spain
Mauro Nicolao, The University of Sheffield, UK
Alexandre Nikolaev, Helsinki Collegium for Advanced Studies, Finland
Marcus Nyström, University of Lund, Sweden
Aurélie Pistono, Ghent University, Belgium
Vassiliki Rentoumi, SKEL, NCSR Demokritos, Greece
Fabien Ringeval, Université Grenoble Alpes, France
Frank Rudzicz, Toronto Rehabilitation Institute and the University of Toronto, Canada
Ineke Schuurman, KU Leuven, Belgium
Kairit Sirts, University of Tartu, Estonia
Charalambos Themistocleous, Johns Hopkins University, Baltimore, USA
Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK
Magda Tsolaki, Aristotle University of Thessaloniki, Greece
Spyridoula Varlokosta, National and Kapodistrian University of Athens, Greece
Yasunori Yamada, IBM Research, Tokyo, Japan
Stelios Zygouris, Aristotle University of Thessaloniki, Greece

Invited Speaker:

Athanasios Tsanas, the Usher Institute, University of Edinburgh, UK.

Table of Contents

An NLP Pipeline as Assisted Transcription Tool for Speech Therapists

Gloria Gagliardi[1], Lorenzo Gregori[2], Andrea Amelio Ravelli[2]
[1]University of Naples "L'Orientale", [2]University of Florence
gloria.gagliardi@gmail.com, lorenzo.gregori@unifi.it, andreaamelio.ravelli@unifi.it

Abstract

This work presents the design of a computer-assisted transcription system for speech-language therapists and an evaluation of its core-module: the NLP pipeline. This pipeline combines a tokenizer, a lemmatizer, a part-of-speech tagger and a spellchecker to perform a semi-automatic annotation of speech transcriptions. The implemented module has been evaluated on a corpus of spoken interaction of children with Developmental Language Disorder (DLD) with the caregiver. Results are promising in automatic error detection (F-measure of 0.547 against a Ground Truth of 0.616) but low in automatic error correction, and confirm the effectiveness within an assisted transcription tool.

Keywords: Pathological Speech Processing, Developmental Language Disorder, PoS Tagging, Lemmatization

1. Introduction

Speech-language assessment and treatment are complex processes. Describing and interpreting children's communication abilities entail the integration of a variety of information, gathered in the evaluation process (e.g. case history, review of sensory-motor and cognitive status, standardized and non-standardized measures of verbal and non-verbal language) (American Speech Language Hearing Association, 2004). The analysis of spontaneous and semi-spontaneous spoken productions of young patients is one of the essential elements for the formulation of logopedic balance, to ascertain the type, factor(s), and severity of the speech-language disorders (such as Speech Sound Disorder, Developmental Language Disorder and Social pragmatic Communication Disorder (American Psychiatric Association, 2013)), and to evaluate the expected habilitation or rehabilitation potential to set functional goals.

In the common practice, documentation of linguistic competence usually includes a portfolio of the child communication samples, e.g. transcript of audio or video-recorded interactions. To date, the collection and analysis of these data are very time consuming: as a matter of fact, Italian therapists manually transcribe the samples using phonetic alphabet (i.e. IPA, International Phonetic Alphabet), and this work is usually performed on "paper". As a result, all the quantitative information which is needed for the evaluation (e.g. number/type of phonemic errors, number of tokens and lemmas, Mean Length of Utterance - MLU) is also empirically computed, representing a huge waste of time and resources.

1.1. Automatic annotation of pathological spoken language: a new challenge for the NLP community

Part-of-speech (PoS) tagging and lemmatization represent important preprocessing steps in Natural Language Processing: they are almost indispensable for the exploitation of corpus data and, since PoS tags are an essential input for most syntactic parsers, the accuracy of their annotation transitively worsens all the subsequent downstream higher level processing tasks (e.g. relation extraction) (Fan et al., 2011; Ferraro et al., 2013).

POS-tagging is actually considered a solved task, since state-of-the-art taggers' accuracy is around 97%–98% for English (Manning, 2011) and, nowadays, tools showing comparable outcomes are available for most languages, including Italian (Tamburini, 2007; Attardi and Simi, 2009; Tamburini, 2013).

As stated by (Giesbrecht and Evert, 2009) this means that, on average, every sentence contains a tagging error, but the accuracy of the system is close to the level of agreement between human annotators, and thus to the upper limit that can be expected from an automatic tool. This high accuracy is mostly attributable to the large amounts of tagged corpora, and the rapid progress in the study of corpus-based computational linguistics.

However, the state-of-the-art POS-taggers trained on written corpora do not provide satisfactory results if applied to spontaneous and semi-spontaneous spoken language (Uchimoto et al., 2002; Panunzi et al., 2004). Essentially, it is due to some peculiarities of the "oral medium", namely freest word order, repetitions and fragmentation phenomena like false starts and interruptions.

Furthermore, PoS tagging and lemmatization tasks on speech corpora have not been tackled yet by the EVALITA periodic evaluation campaigns of NLP tools for the Italian language.[1]

Clearly, this lack in NLP for spoken Italian also affects the automatic analysis of children's verbal productions and adult pathological language (e.g. aphasic speech).

The limited availability of data remains a stumbling block to reach state-of-the-art performances of NLP tools in the clinical domain. However, the number of computational applications is growing rapidly in the medical field: NLP techniques have been applied to the analysis of patients' written and spoken texts, revealing latent patterns and regularities of their verbal productions, and thus acting as "digital biomarkers" (i.e. objective, quantifiable behavioral data which can be collected and measured through digital device, allowing for low-cost pathology detection and classification).

[1] http://www.evalita.it/

2. Towards a computer-assisted transcription tool

Within the NLP tools for clinical application, we designed a system to support speech-language therapists in the error analysis of spoken productions. We aim at facing this issue by proposing an NLP pipeline for the assisted transcription and automatic analysis of speech recordings collected from Italian typical/atypical developing children. To the best of our knowledge, no previous study addressed this issue up till now for the Italian language.

In our intentions, the tool should support the speech-language therapists during all the phases, reducing their work burden. As a matter of fact, a simple but effective pipeline will allow the speech-language therapist to transcribe and automatically analyse spoken texts; the workflow can be summarised as follows:

1. *Transcription*: the user digitally transcribes the recorded samples, using the SAMPA phonetic alphabet (Wells, 1997).
2. *SAMPA to orthographic transcription converter*: the system converts phonetic transcriptions to regular Italian graphemes, so to be processed by an NLP pipeline.
3. *First automatic annotation*: tokenization, PoS Tagging and lemmatization of raw texts.
4. *Assisted transcription/correction module*: the system highlights "idiosyncratic words", suggesting possible "corrections" by means of a spellchecker (e.g. il lubo > il lupo, en. *'the wolf'*).
5. *Manual correction of misspelled words*.
6. *Final automatic annotation*: PoS tagging and lemmatization of "normalized" texts.
7. *Statistics and IPA phonetic transcription generation*.

The full procedure requires limited user training. Italian therapists are usually reluctant to digitally transcribe, due to discomfort and concerns about the IPA keyboard. This difficulty can be easily overtaken using the SAMPA chart (Speech Assessment Methods Phonetic Alphabet), which is a machine-readable phonetic alphabet (Table 1).

As a matter of fact, the mapping of phonology into orthography is quite transparent and regular for Italian, where differences are limited to few phonemes. Therefore, phonetic and orthographic transcriptions are almost equivalent from a practical point of view. The initial effort is balanced out by the time saved in the analysis stage: after the final annotation, the system can quickly extract statistics at the phonological, lexical, and morpho-syntactic level, by comparing the raw transcription with the normalized one. For example, the following phonological processes can be easily identified:

- Consonant cluster reduction

  ['kwesto]> ['kwetto] (*'this'*), ['skappa]> ['kappa] (*'runs away'*)

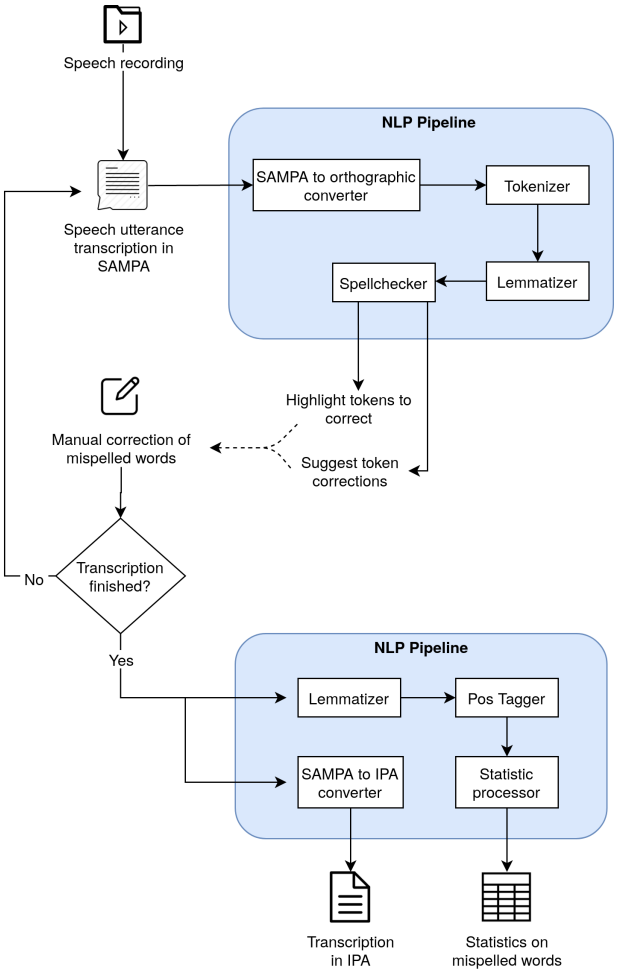

Figure 1: Full computer-assisted transcription pipeline.

Description	IPA	SAMPA
bilabial plosive	p b	p b
alveolar plosive	t d	t d
velar plosive	k ɡ	k g
bilabial nasal	m	m
alveolar nasal	n	n
palatal nasal	ɲ	J
labio-dental fricative	f v	f v
alveolar fricative	s z	s z
palato-alveolar fricative	ʃ ʒ	S Z
alveolar affricate	ts dz	ts dz
palato-alveolar affricate	tʃ dʒ	tS dZ
alveolar trill	r	r
alveolar lateral	l	l
palatal lateral	ʎ	L
approximant	j w	j w
vowels	a ɛ e i ɔ o u	a E e i O o u

Table 1: IPA and SAMPA phonetic alphabets.

- Consonant voicing

  ['lupo]>['lubo] (*'wolf'*)

Classical measures of lexical richness (e.g. Type/Token ratio) and syntactic development (e.g. MLU) can also be

automatically computed, lightening the workload. The pipeline can also generate an IPA transcription, which can be inserted in the patient's portfolio, as requested by national good practice.

This paper presents the core module of the aforementioned pipeline (Figure 1), focusing on the ability to identify misspelled words, to suggest correction candidates and to automatically analyse transcriptions of pathological speech.

3. Material

To test the effectiveness of the pipeline, we rely on a small corpus of transcription of spontaneous speech interaction between infants and caregivers. This resource was designed to provide a first picture of narrative discourses produced by Italian monolingual preschoolers with Developmental Language Disorder (DLD) in comparison with typical peers matched by age.

DLD (previously known as Specific Language Impairment or SLI) is a neurodevelopmental disability which affects linguistic and communicative competence (American Psychiatric Association, 2013; Bishop et al., 2017): it is the most frequent developmental disorder in childhood, with an estimated overall prevalence in pre-school-aged children of about 7% (Tomblin et al., 1997; Johnson et al., 1999). It can selectively compromise all speech and language domains, affecting both language production and comprehension. A diagnosis of DLD should be stated (Bishop et al., 2017) for children showing a lower linguistic competence in comparison with the pairs; this verbal difficulty must affect patients' everyday functioning and is unlikely to resolve by five years of age; in addition, it is not associated with a known cognitive, neurological or sensory-motor differentiating condition, depicting a more complex pattern of impairments, (e.g. brain injury, acquired epileptic aphasia in childhood, cerebral palsy, oral language limitations associated with sensorineural hearing loss as well as genetic conditions such as the Down syndrome).

To build our corpus, sixteen monolingual infants (13 M; 3 F) ranging in age from 4;2 to 5;4 (mean = 4;7) were enrolled. The sample was composed of a Control Group (CG) and a DLD Group, matched by age. The CG included eight participants (5 M; 3 F) without speech, language, hearing or cognitive impairments. The DLD group included eight male children who met the criteria for DLD with expressive deficits (American Psychiatric Association, 2013), recruited through the AUSL Toscana Centro. The diagnosis has been established according to national and international guidelines by expert clinicians, based on anamnestic data, clinical observation and standardized testing. Participants underwent a complete language evaluation, but particular attention has been paid to the assessment of children's comprehension profile: all subjects performed within the normal range on the test of receptive vocabulary (TNL, Test Neuropsicologico Lessicale per l'età evolutiva (Cossu, 2013)), morpho-syntactic comprehension (TCGB, Test di Comprensione Grammaticale per Bambini (Chilosi and Cipriani, 2006) and PVCL, Prove di Valutazione della Comprensione Linguistica (Rustioni and Lancaster, 2007)) and listening comprehension (TOR, Test di Comprensione del Testo Orale 3-8 anni (Levorato and Roch, 2007)); therefore, expressive language problems occur essentially in isolation.

The corpus is composed by caregiver-child spontaneous speech interactions (duration: min. 3'51" - max. 23'53"), for a total of 1h57'41" transcribed audio-visual material. Oral production was elicited through three different tasks: the norm-referenced Bus Story Test (I-BST) (Renfrew, C.E., 2015; Cipriani et al., 2012; Mozzanica et al., 2016), and two semi-spontaneous retelling assessments, exploiting the renowned story Three Little Pigs (3LP), and a brand new short film called Little Polar Bear (LPB). While the I-BST examines story retelling with a colored picture support, the unnormed tests elicit children's verbalizations through a paper book and a tablet respectively. During the 3LP task, children were asked to retell the renowned story using the pictures as prompts while flipping through the pages; in contrast, the LPB task was administered showing the video (around 100 seconds) to the child who was then requested to recount the plot while following the scrolling images without sound. None of the children knew the three stories. The trials were administered in a single test session of varying duration (∼30 minutes).

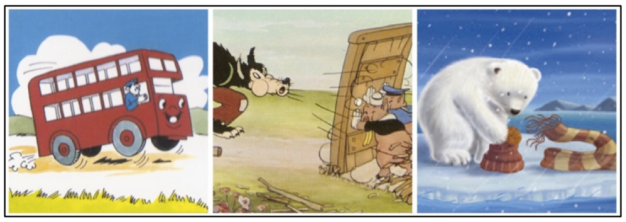

Figure 2: The proposed tasks. From the left: "Bus Story Test", "Three Little Pigs" and "Little Polar Bear".

The tasks were recorded using a tablet placed in front of the subject. Data were transcribed using ELAN (Wittenburg et al., 2006).[2] Furthermore, transcriptions are also compliant with the L-AcT format (Cresti and Moneglia, 2018), a version of the standardized CHAT format (MacWhinney, 2000) enriched with the tagging of prosodic parsing. We chose the *utterance* as the reference unit in the speech continuum, defined as the counterpart of a speech act, namely 'the minimal linguistic entity that can be pragmatically interpreted' (Austin, 1962; Cresti and Moneglia, 2018)). Utterances are demarcated by prosody in the speech flow, therefore the identification of their boundaries is achieved through the detection of "prosodic breaks". The identification of breaks reaches high inter-rater agreement in annotation, also among non-expert annotators (Cohen's kappa for Italian around 0.8; (Danieli et al., 2004)), thus being a highly reliable chunking method.

4. The NLP pipeline in detail

The designed pipeline takes as input the text of the transcription of the session, and gives as output 2 objects: tran-

[2] All parents gave their consent to data recording, transcribing and processing.

scription in IPA characters, and detailed statistics of errors per part-of-speech and lemma.

Starting from the transcription of the session, the first step in the pipeline is the conversion from SAMPA to orthographic text, through a simple set of re-writing rules. Moreover, L-AcT specific tags and annotations (e.g. rephrasing, false starts, etc.) are removed, and the speaker's turns are stored as distinct strings to be processed individually. In this way, it is possible to focus all the analysis exclusively on the child turns.

Then, each turn is tokenized and lemmatized with TreeTagger (Schmid, 1994), and all tokens are analysed by a spellcheck module. We used *pyspellchecker*,[3] a Python module that implements a Levenshtein Distance algorithm (Levenshtein, 1966) to find all possible permutations within an edit distance of 2 characters from each misspelled word. It then compares all permutations (character insertions, deletions, replacements, and transpositions) to known words in a word frequency list. As reference dictionary for the spellchecker module we used an Italian list of 50k words from the *WordFrequency Project*[4], that has been extracted from the OpenSubtitles multilingual corpus[5] (Lison and Tiedemann, 2016).

The word that is found more often in the frequency list is more likely the correct result, and it is proposed as a substitution for the entry. At this point, the human annotator can chose to accept the proposed correction, or reject it and manually type the correct word. The index of the misspelled and its correction is stored, to be further used for the error analysis. The edited version of the text is then passed back to TreeTagger to perform lemmatization and POS-tagging. We perform lemmatization twice because misspelled words are initially tagged as "unknown", and we make use of the tag shift in the error analysis.

Statistics on misspelled words can be easily obtained by parsing the annotated text. As an example, two fundamental pieces of information for a therapist are the set of wrong pronunciation of the same word and part-of-speech distribution during the speech.

Finally, the whole SAMPA transcript is converted to IPA, similarly to the very first step of the pipeline (SAMPA to orthographic), by following a simple set of re-writing rules, and the complete session is written out as a text file.

DLD Group			
	Child	**CG**	**Total**
Tokens	3367	3840	7207
Words	2191	2639	4830
Unique words	467	433	702
Unique lemmas	296	270	403
Type/token ratio	0.135	0.102	0.083

Table 2: Number of tokens, words and lemmas produced by children and care givers in the DLD Group sessions.

[3] https://pypi.org/project/pyspellchecker/

[4] https://github.com/hermitdave/FrequencyWords

[5] http://opus.nlpl.eu/OpenSubtitles2018.php

Control Group			
	Child	**CG**	**Total**
Tokens	3419	2338	5757
Words	2345	1652	3997
Unique words	514	385	665
Unique lemmas	345	282	431
Type/token ratio	0.147	0.170	0.108

Table 3: Number of tokens, words and lemmas produced by children and care givers in the Control Group sessions.

4.1. Corpus statistics

The automatic annotation through the NLP pipeline allowed us to derive some interesting information about the corpus used for this work. Tables 2 and 3 report the number of tokens, words, lemmas, and type/token ratio of DLD and Control Group sub-corpora. We can see that the number of words produced by children is similar between the two groups (2191 and 2345), attesting a substantial balance in the data.

An interesting result is that no relevant differences emerge between the two groups regarding the type/token ratio: 0.135 for DLD and 0.147 for Control. It derives that the speech of children with language disorder have roughly the same lexical variety than the speech of typical children. Otherwise, a big difference can be observed in caregiver speech (0.102 in DLD and 0.170 in Control), highlighting that caregivers talk is more simplified when addressed to children with language disorder.

PoS	**DLD**	**Control**
Noun	453 (20.68%)	453 (19.50%)
Verb	451 (20.58%)	457 (19.67%)
Conjunction	410 (18.71%)	384 (16.53%)
Article	256 (11.68%)	232 (9.99%)
Preposition	131 (5.98%)	139 (5.98%)
Adjective	108 (4.93%)	86 (3.70%)
Clitic	**90 (4.11%)**	**143 (6.16%)**
Adverb	89 (4.06%)	104 (4.48%)
Pronoun	**61 (2.78%)**	**105 (4.52%)**
Articulated Prep.	44 (2.01%)	66 (2.84%)
Determiner	31 (1.41%)	48 (2.07%)
Auxiliary verb	23 (1.05%)	28 (1.21%)
Negation	20 (0.91%)	39 (1.68%)
Word "che"	17 (0.78%)	28 (1.21%)
WH Word	5 (0.23%)	2 (0.09%)
Proper Noun	2 (0.09%)	4 (0.17%)
Number	0 (0.00%)	5 (0.22%)

Table 4: Part-of-speech distribution in children speech (in DLD and Control Groups).

Finally, when looking at the part-of-speech distribution of children in the two groups (Table 4) we could not find huge differences, but a notable gap can be observed in the production of clitics and pronouns, where numbers are lower in DLD Group (χ-squared test with p-value < 0.001). This

seems to confirm and enrich known data about clitic productions in Italian impaired children (Bortolini et al., 2006; Guasti et al., 2016), even if further analyses are needed to support this argument.

5. Evaluation

Table 5 reports the output of the error analysis performed within the NLP pipeline. POS_unknown refers to the lemmas not recognized by the POS-tagger, while Spellcheck stands for the words reported by the spellchecker. It is possible to notice a slight difference between the phenomena highlighted by the two methods.

	DLD	Control
POS_unknown	90 (4.11%)	40 (1.71%)
Spellcheck	84 (3.83%)	55 (2.35%)

Table 5: Number of words tagged as "unknown" by the POS-tagger and marked by the spellchecker. Percentages are reported with respect to the total number of words in each group.

To evaluate the results of error identification and automatic correction tasks, we built a gold standard through manual annotation of the children turns in the whole corpus. Each misspelled word were marked and annotated with the correct version. Data reported in Table 6 show that, as expected, the DLD Group has a double rate of misspelled words than the Control Group: 4.11% of the total produced words in DLD are misspelled against 2% in Control. Moreover, it is important to highlight that a significant number of misspelled words are not recognized by the human annotator which marked them with "unknown" during the manual check. In total, there are 8 words in the DLD Group and 10 in the Control group, for a total of 13,14% of misspelled words.

	DLD	Control
Manual corr. (MC)	90 (4.11% w.)	47 (2.00%)
MC unclassified	8 (8.89% MC)	10 (21.28%)

Table 6: Number of manual corrections in the gold standard (total and unknown words).

Accuracy in both error detection and correction is reported in Table 7. For the error detection task we reported the number of manual corrections matching with the "unknown" tag of the lemmatizer (Lem) and the number of manual corrections matching the words marked by the spellchecker (SC). Automatic correction task is performed by the spellchecker only and the numbers regard automatic corrections matching with manual corrections. Precision, Recall and F-measure are computed for both tasks and reported in Table 8.

It is important to highlight that the proposed system is not able to identify any of the cases in which the misspelled word is still a word form that exists in the language. These cases are frequent in Italian, especially with short words, like articles or prepositions, where it is likely that deletion

	DLD	Control
Err. detection (Lem)	55 (61.11%)	18 (38.30%)
Err. detection (SC)	48 (53.33%)	18 (38.30%)
Err. correction (SC)	27 (32.14%)	4 (7.27%)

Table 7: Numbers and percentages of misspelled words properly detected by lemmatizer (Lem) and spellchecker (SC), and properly corrected ones by the spellchecker.

	Pr	Rec	Fm
Error detection (Lem)	0.562	0.533	0.547
Error detection (SC)	0.475	0.482	0.478
Error detection (GT)	0.681	0.562	0.616
Error correction (SC)	0.223	0.304	0.257

Table 8: Precision, Recall and F-measure of the error detection task for lemmatizer (Lem), spellchecker (SC) and Ground Truth (GT), and of the error correction task for spellchecker.

or substitution of a single phoneme produce a proper word (e.g. *il* > *i*; *del* > *dei*). For this reason, the maximum Recall that our system can reach in error identification task cannot be very high: with the given dataset, considering only errors that produce impossible words, we obtained a ground truth Recall of 0.562. Table 8 shows that there is a low margin of improvement. Conversely, the Precision of the system is deeply affected by those lexical productions that are specific of spoken language, like interjections, vocalizations and filled pauses (e.g. *ehh*, *mah*, *mmm*), which are wrongly marked as errors. Considering these expressions in our dataset as constrained false positive, we obtained a ground truth Precision of 0.681.

While a substantial Recall improvement is not possible with the given system - because it would require additional NLP modules of language understanding - Precision in error detection could be improved a lot, by upgrading the pipeline with NLP tools (spellchecker and lemmatizer) trained on spoken corpora.

As stated before, some errors cannot be satisfactorily managed by the pipeline. As an example, there are some phonological processes that are typical in children linguistic development which result in real words (e.g. ['tʃuffo] > ['tuffo], *stopping*, en. *'lock of hair'* > *'dive'*; [ba'nana] > ['nana], *weak syllable deletion*, en. *'banana'* > *'dwarf'*) and neologisms like 'peciano', 'selfia' or the portmanteau 'fangua' (coined by blending 'fango' and 'acqua', en. *'mud'-'water'*). These phenomena are not understandable by the therapist outside their linguistic and extra-linguistic contexts. On the contrary, simple heuristics can be incorporated into the pipeline to manage high-frequency articulation or phonological error patterns that characterised typical and atypical developmental trajectories. For example, the already mentioned cluster reduction (e.g. ['kwesto] > ['kwetto], en. *'this'*, or ['skappa] > ['kappa], en. *'ø run away'*), prevocalic consonant voicing (e.g. ['lupo] > ['lubo], en. *'wolf'*) or deaffrication (['gottʃe] > ['gosse], en. *'drops'*).

By considering these data and their analysis, we can derive that a semi-automatic system of computer-assisted translation, as proposed in this paper, appears to be more suitable than a fully automatic one, that provides an automatic annotation of transcripts. In fact, results of error detection are promising and can be fruitfully exploited to highlight misspelled words, while accuracy on automatic correction is low and definitely not reliable to replace manual annotation. However, proposed corrections can be very useful to save time during the transcription, showing a set of possible correction options that can be selected by the annotator. To this aim, a simple caching system of annotated data would bring a strong improvement to the spellchecker, given that, phonological errors tend to be recurrent: in our test corpus, 40.33% of the pairs [correct word, misspelled word] occur more than once.

6. Conclusions and future work

This work discussed the application of an NLP pipeline within a computer-assisted transcription system. The system architecture foresees a SAMPA transcription of pathological speech and aims at helping speech therapists to annotate misspelled words, to produce useful statistics on errors in words production, and to generate text in IPA. The core module of the system was developed and analyzed through a spoken corpus of children with Developmental Language Disorder. The tasks considered are automatic detection and automatic correction of misspelled words. The evaluation highlights an average accuracy on error detection and a low accuracy on error correction. However the results appear to be relevant for the proposed application. It is important to notice that a naif spellchecker module was implemented, thus more sophisticated systems may be able to improve also error correction results. It is important to point out that the lack of large annotated speech corpora for Italian (and in particular for first language acquisition) is the main obstacle to a more effective system. In fact, many of the problems highlighted in this paper would be correctly handled by NLP tools specifically trained on spoken Italian. The presented analysis represents the first step in the construction of a full transcription tool that will be developed as an editor for speech therapists (in the form of a standalone software or ELAN plugin).

7. Acknowledgements

The authors are deeply grateful to Francesca Beraldi and Milvia Innocenti who collected the clinical data. The precious help of Annalisa Raffone is also is also acknowledged.

8. Author contribution

As far as academic requirements are concerned, sections 1 and 3 were authored by Gloria Gagliardi, who also performed the manual annotation for the pipeline evaluation; Lorenzo Gregori wrote sections 2 and 5 and computed the statistics; Andrea Amelio Ravelli takes official responsibility for section 4 and implemented the NLP pipeline. All authors read and gave final approval for submission. The usual disclaimers apply.

9. Bibliographical References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. American Psychiatric Association, Washington, DC.

American Speech Language Hearing Association. (2004). *Preferred Practice Patterns for the Profession of Speech-Language Pathology*.

Attardi, G. and Simi, M. (2009). Overview of the evalita 2009 part-of-speech tagging task. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence, 12th December 2009, Reggio Emilia, Italy*.

Austin, J. (1962). *How to Do Things with Words*. Clarendon Press, Oxford.

Bishop, D., Snowling, M., Thompson, P., Greenhalgh, T., and THE CATALISE-2 CONSORTIUM. (2017). Phase 2 of catalise: a multinational and multidisciplinary delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry*, 58(10):1068–1080.

Bortolini, U., Arfé, B., Caselli, C. M., Degasperi, L., Deevy, P., and Leonard, L. B. (2006). Clinical markers for specific language impairment in italian: the contribution of clitics and non-word repetition. *International Journal of Language & Communication Disorders*, 41(6):695–712.

Chilosi, A. and Cipriani, P. (2006). *TCGB. Test di comprensione grammaticale per bambini*. Edizioni Del Cerro, Tirrenia.

Cipriani, P., Salvadorini, R., and Zarmati, G. (2012). *Bus Story Test. Test di valutazione delle abilità narrative*. Edizioni La Favelliana, Milano.

Cossu, G. (2013). *TNL. Test neuropsicologico lessicale per l'età evolutiva*. Hogrefe Editore, Firenze.

Cresti, E. and Moneglia, M. (2018). Chapter 13. the illocutionary basis of information structure: The language into act theory (l-act). pages 360–402.

Danieli, M., Garrido, M., Moneglia, M., Panizza, A., Quazza, S., and Swerts, M. (2004). Evaluation of Consensus on the Annotation of Prosodic Breaks in the Romance Corpus of Spontaneous Speech "C-ORAL-ROM". In Maria Teresa Lino, et al., editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1513–1516, Paris. ELRA.

Fan, J.-W., Prasad, R., Yabut, R. M., Loomis, R. M., Zisook, D., Mattison, J. E., and Huang, Y. (2011). Part-of-speech tagging for clinical text: wall or bridge between institutions? *AMIA Annual Symposium*, pages 382–91.

Ferraro, J. P., Daumé, H., DuVall, S. L., Chapman, W. W., Harkema, H., and Haug, P. J. (2013). Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association: JAMIA*, 20(5):931–9.

Giesbrecht, E. and Evert, S. (2009). Is part-of-speech tagging a solved task? an evaluation of pos taggers for the

german web as corpus. In I. Alegria, et al., editors, *Proceedings of the 5th Web as Corpus Workshop (WAC5)*.

Guasti, M. T., Palma, S., Genovese, E., Stagi, P., Saladini, G., and Arosio, F. (2016). The production of direct object clitics in pre-school– and primary school–aged children with specific language impairments. *Clinical Linguistics & Phonetics*, 30(9):663–678. PMID: 27285056.

Johnson, C., Beitchman, J., Young, A., Escobar, M., Atkinson, L., Wilson, B., Brownlie, E., Douglas, L., Taback, N., Lam, I., and Wang, M. (1999). Fourteen-year follow-up of children with and without speech/language impairments: speech/language stability and outcomes. *Journal of Speech, Language, and Hearing Research*, 42(3):744–760.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Levorato, M. and Roch, M. (2007). *TOR. Test di Comprensione del Testo Orale - 3-8 anni*. Giunti O.S., Firenze.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Lowrence Erlbaum Associates, Mahwah.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I*, CICLing'11, pages 171–189. Springer-Verlag.

Mozzanica, F., Salvadorini, R., Sai, E., Pozzoli, U., Maruzzi, P., Scarponi, L., Barillari, M., Spada, E., Ambrogi, F., and Schindler, A. (2016). Reliability, validity and normative data of the italian version of the bus story test. *International Journal of Pediatric Otorhinolaryngology*, 89:17–24.

Panunzi, A., Picchi, E., and Moneglia, M. (2004). Using pitagger for lemmatization and pos tagging of a spontaneous speech corpus: C-oral-rom italian. pages 563–566. ELRA, Paris.

Renfrew, C.E. (2015). *Bus Story Test. A test of narrative speech*. Speechmark, Milton Keynes (UK).

Rustioni, D. and Lancaster, M. (2007). *PVCL Valutazione della Comprensione Linguistica*. Giunti O.S., Firenze.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Tamburini, F. (2007). Evalita 2007: The part-of-speech tagging task. *Intelligenza Artificiale*, IV(2):4–7.

Tamburini, F. (2013). The lemmatisation task at the evalita 2011 evaluation campaign. In B. Magnini, et al., editors, *Evaluation of Natural Language and Speech Tools for Italian*, pages 230–238. Springer.

Tomblin, J., Smith, E., and Zhang, X. (1997). Epidemiology of specific language impairment: Prenatal and perinatal risk factors. *Journal of Communication Disorders*, 30(4):325 – 344.

Uchimoto, K., Nobata, C., Yamada, A., Sekine, S., and Isahara, H. (2002). Morphological analysis of the spontaneous speech corpus. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.

Wells, J. C. (1997). Sampa computer readable phonetic alphabet. *Handbook of standards and resources for spoken language systems*, 4.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. pages 1556–1559.