



UNIVERSITÀ DEGLI STUDI DI SALERNO

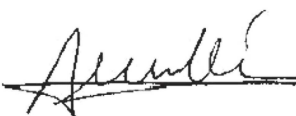
Dipartimento di Scienze Politiche e della Comunicazione

Corso di Dottorato in Scienze del Linguaggio, della Società,
della Politica e dell'Educazione

Tesi di Dottorato

Italian Multimodal Corpus:

Verbal and Non-Verbal Communication in Political Domain

Tutor: 

Prof. Annibale Elia

Co-Tutor:

Dott.ssa Sara Tonelli



Coordinatore:

Prof. Filippo Fimiani



Candidato:

Daniela Trotta

Matr. 8801400078



Contents

1	Introduction	1
1.1	Research Questions and Contribution	3
2	Multimodal Corpora	5
2.1	Political-Domain Corpora	7
3	PoliModalCorpus 2.0.	11
3.1	Description of the PoliModal Corpus	13
3.1.1	Annotation Scheme	14
	Metadata annotation	15
	Utterances and speaking turns	16
	Pausing	17
	Vocal	18
	False starts, repetitions and truncated words	19
	Overlap	21
	Inter-annotator agreement	22
3.2	Corpus Analysis	23
3.2.1	Statistics of Non-Verbal Traits	24
3.2.2	Political Orientation and Language Use	25
3.2.3	Relation Between Verbal and Non-Verbal Traits	29
3.3	Multimodal Annotation	32
3.4	Corpus Statistics and Discussion	38

4	Co-Gesture Analysis	43
4.1	Gesturing with hands	46
4.2	Coding co-speech gesture	49
4.3	Which type of verb do hand movements accompany most frequently?	51
4.4	Is the Lexical Retrieval hypothesis confirmed?	59
4.4.1	Background to Lexical Retrieval Hypothesis	60
4.4.2	Verification of the Lexical Retrieval Hypothesis	61
5	Results and Discussion	63
5.1	Is the gesture-speech relationship influenced by linguistic variables?	67
5.2	Semantic gesture-speech relationship	69
5.2.1	Political party influences the type of gestures: the one-way ANOVA test	75
6	Conclusions, future work and perspectives in computational linguistics	77
6.0.1	Machine learning algorithms and multimodal corpora	79
	Bibliography	83

List of Figures

3.1	Raw text	15
3.2	Metadata annotation	16
3.3	Annotation of utterances	17
3.4	Pausing annotation	18
3.5	Vocal annotation	19
3.6	False Start annotation	20
3.7	Annotation of repetition	21
3.8	Truncation annotation	21
3.9	Overlap annotation	22
3.10	Distribution of traits per political party (avg. number of occurrences per turn)	25
3.11	Use of nouns, verbs, adjectives and adverbs for each politician (% over all content words)	27
3.12	Avg. nouns per political party	30
3.13	Example of facial display: frowning.	36
3.14	Example of hand gesture: double-handed.	38
3.15	Example of body posture: sideways.	39
4.1	Annotation extract in xml	50
4.2	Traditional formula of Mutual Information	61
4.3	Weighted Mutual Information formula	62
5.1	Avg. hand movements per political party	64
5.2	WMI values for each tag divided by interviewee	65

List of Tables

3.1	Corpus statistics related to the 13 interviews included in our study	23
3.2	List of gestures, following the list described in (Allwood et al., 2007). The presence of bold means that the gesture has been found in our dataset.	34
3.3	Corpus content: turns per speaker and total duration	40
3.4	Statistics on annotated information comparing number of occurrences and average duration in milliseconds.	41
4.1	Absolute frequencies of verbal modes	54
5.1	Normalized frequencies of the tags for each interview	63
5.2	Normalized values of hand movements, TTR, and lexical density for each interviewee	68
5.3	Values of hand movements, TTR, and lexical density for each political party	68
5.4	Frequency of the type of gestures produced by each interviewee	70
5.5	Frequency of the type of gestures for each political party	71
5.6	ANOVA test results	76

Chapter 1

Introduction

Speaker gestures are semantically co-expressive with speech and serve different pragmatic functions to accompany oral modality. Therefore, gestures are an inseparable part of the language system: they may add clarity to discourse, can be employed to facilitate lexical retrieval and retain a turn in conversations, assist in verbalizing semantic content and facilitate speakers in coming up with the words they intend to say. This aspect is particularly relevant in political discourse, where speakers try to apply communication strategies that are both clear and persuasive using verbal and non-verbal cues.

This dissertation aims to analyze the co-speech gestures of several Italian politicians during face-to-face interviews using a multimodal linguistic approach.

By ‘multimodal’ we mean that the corpus is composed by audio-video recordings of interviews broadcast on TV with an orthographic transcription, which aims at the transposition of speech into the standard of the writing system, assuming as a conventional reference entity the graphic word. In our case, the transcriptions are annotated with information not only about the linguistic structure of the utterances but also about non-verbal expressions¹.

¹According to (Allwood, 2008): “The basic reason for collecting multimodal corpora is that they provide material for more complete studies of ‘interactive face-to-face sharing

The work first introduces the corpus created: PoliModal corpus (Trotta et al., 2019; Trotta et al., 2020), containing the transcripts of 56 TV face-to-face interviews of 14 hours, taken from the Italian political talk show “In mezz’ora in più” (for a total of 100,870 tokens) that has been manually annotated with information about metadata (i.e. tools used for the transcription, link to the interview etc.), pauses (used to mark a pause either between or within utterances), vocal expressions (marking non-lexical expressions such as coughs and semi-lexical expressions such as primary interjections), deletions (false starts, repetitions and truncated words), overlaps and facial displays, hand gestures and body posture ².

Then, the annotation scheme and the results of a series of statistical analyses aimed at understanding the relationship between annotated multimodal traits and language complexity are described in detail and testing the validity of existing studies on political orientation and language use.

Finally, after the presentation of an additional semantic annotation layer related to the function assumed by hand movements, the relationship between them and other information layers such as a political party or non-lexical and semi-lexical tags is investigated.

Concerning gesture speech relationship, the results obtained suggest that hand movements are mainly used with integrative and complementary functions. So, the information provided by such gestures adds precision and emphasis to spoken information. Its, also show that party affiliation does not significantly influence the gesture-speech relationship.

Furthermore - testing the lexical retrieval hypothesis by calculating the association between the hand movements produced by each respondent and discourse disfluencies using weighted mutual information - it is and construction of meaning and understanding’ which is what language and communication are all about.”

²The corpus is freely available for research purpose at the link <https://github.com/dhfbk/InMezzoraDataset>

shown that hand movements tend to co-occur with full pauses (i.e., repetition) and empty pauses (i.e., pause) and more frequently with interjections (semi-lexical tags), suggesting that gesticulation may represent an attempt at lexical retrieval.

1.1 Research Questions and Contribution

The dissertation aims to demonstrate the great potential of using multi-modal corpora annotated on multiple levels, specifically answering the following research questions:

1. Which type of verb do hand movements accompany most frequently?
2. Since the corpus used as a case study presents an annotation of so-called "speech constants" (Voghera, 2001) (i.e. pauses, interjections, false starts, repetitions, truncations), is the Lexical Retrieval hypothesis confirmed or are gestures used in correlation with other and different constants of speech? Note that the Lexical Retrieval hypothesis assumes that (a) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978), and (b) that the inability to gesture can cause verbal disfluencies (Dobrogaev, 1929).
3. Is the gesture-speech relationship influenced by linguistic variables such as the complexity of the language in terms of type-token ratio and lexical density?
4. What are the semantic patterns of gesture-speech relationship? Does political party affiliation influence this relationship?

The main contribution of this research work is to be found in the release of the annotated resource (freely available on the Github, Clarin and Accademia della Crusca databases³, which therefore can be used by scholars not only in the field of linguistics (e.g., political science) for further observations and in the field of computational linguistics as a basis for the study of communicative behavior in the political sphere through the use of machine learning algorithms in order to improve current *Argument Mining* methodologies in the identification of argumentative structures in the text. This aspect will be among others discussed more explicitly in the conclusions.

³The resource created is to date freely accessible on the following platforms of national and international prominence:

- Github: <https://github.com/dhfbk/InMezzoraDataset/blob/master/README.md>
- CLARIN infrastructure: <https://www.clarin.eu/resource-families/multimodal-corpora>
- ILC4CLARIN, European Language Resource Infrastructure for the Humanities and Social Sciences: <http://hdl.handle.net/20.500.11752/OPEN-534>
- Accademia della Crusca, Institutional, legal and administrative Italian subsection: <https://accademiadellacrusca.it/it/contenuti/banche-dati-corpora-e-archivi-testuali/6228>

Chapter 2

Multimodal Corpora

The concept of a multimodal corpus has been defined by (Allwood, 2008) in terms of an annotated collection of “language and communication-related material drawing on more than one modality”. Multimodal corpora (or multimedia corpora as they are often defined in the Italian literature) are used especially for pragmatic research purposes (i.e. in studies on proxemic correlates of spoken language or on the bodily manifestation of emotions), in which the starting sessions consist of videos that are transcribed and annotated (Cresti and Panunzi, 2013). According to (Allwood, 2001), research questions can be divided into three major areas:

1. *human-human face-to-face communication*: the nature of communicative gestures, multimodal communication in different national/ethnic cultures, communication and consciousness/awareness, etc;
2. *media of communication*: multimodality in writing, multimodality in songs and music, etc.;
3. *applications*: better modes of multimodal human-computer communication, better modes of multimodal distance teaching/instruction, etc.;

In addition, multimodal corpora can be useful resources in the development of various computer-based applications, supporting or extending

our ability to communicate, with regard to: modes of multimodal human-computer communication, better computer support for multimodal human-human communication, modes of multimodal communication for persons who are physically challenged (handicapped), modes of multimodal presentation of information from databases (for example for information extraction or for summarization), better multimodal modes of translation and interpretation, modes of multimodal distance language teaching (including gestures), better multimodal modes of buying and selling (over the internet, object presentation in shops, etc.), computerized multimodal corpora can, of course, also be useful outside of the areas of computer-based applications. In general, they can provide a basis for studying any communicative behavior in order to fine-tune and improve that behavior.

However, these resources – probably due to the difficulty of construction – in Italy are difficult to find and consult, in fact between the 286 multimodal resources certified for all the languages by the LRE map¹ only one is in Italian, IMAGACT, a corpus-based ontology of action concepts, derived from English and Italian spontaneous speech (Moneglia et al., 2014; Bartolini et al., 2014). So this language is not well represented.

Notice that a particular type of multimodal product is television interviews. According to (Vignozzi, 2019) they are inherently a multimodal and multi semiotic text, in which meaning is created through the intersection of visual elements, verbal language, gestures, and other semiotic cues.

In television interviews, non-verbal aspects are essential, especially in high emotional involvement and high-stakes contexts. For these reasons, one of the domains more suitable for this kind of analysis is the political

¹LRE map (Language Resources and Evaluation) is a freely accessible large database on resources dedicated to Natural language processing. The original feature of LRE Map is that the records are collected during the submission of different major Natural language processing conferences. The records are then cleaned and gathered into a global database called “LRE Map” (Calzolari et al., 2012). The map is freely available from the site <https://lremap.elra.info/>

one (Seiter and Weger Jr, 2020). In recent years, a remarkably successful research line has focused on analyzing gestures the speaker uses to discredit the opponent. These aspects have been the subject of various studies, even in the Italian language (D’Errico, Poggi, and Vincze, 2010).

Concerning Italian language, some corpora have been made available recently, the largest one includes around 3,000 public documents by Alcide De Gasperi (Tonelli, Sprugnoli, and Moretti, 2019) that has been mainly used to study the evolution of political language over time (Menini et al., 2020). All the corpora cited above are monomodal and none of them considers gestural traits. Indeed, corpora that include only one modality have a long tradition in linguistics. According to (Lin, 2017, p. 157) “the construction and use of multimodal corpora is still in its relative infancy. Despite this, work using multimodal corpora has already proven invaluable for answering various linguistic research questions that are otherwise difficult to consider”.

It should be noted that none of the existing Italian-language resources present a systematic annotation of gestures. As far as it is known, no studies have focused on the presence and behavior of co-gestural patterns for this language.

2.1 Political-Domain Corpora

In recent years, political language has received increasing attention, especially in the Anglo-Saxon and American world, where it is possible to have free access to speech transcriptions from government portals and personal foundation websites, e.g. White House portal, William J. Clinton Foundation, Margaret Thatcher Foundation. This has fostered research on political and media communication and persuasion strategies (Guerini, Straparava, and Stock, 2008; Esposito et al., 2015). At the same time, the technological advancements witnessed in the last twenty years have provided

linguists with better tools for recording, storing, and querying multiple forms of digital records, allowing both verbal and non-verbal elements of language production to be tracked simultaneously. This has provided the foundations for the recent, increasing interest in multimodal corpus linguistics (Knight, 2011).

This trend is highlighted by a large number of workshops and events on the topic, attracting an international, interdisciplinary audience, see for example the GESPIN (Gesture and Speech in Interaction) conference series held in Poznan´ (2009) and Bielefeld (2011), Audio-Visual Speech Processing Workshops (AVSP), LREC Workshops on Multimodal Corpora and the ISGS (International Society for Gesture Studies) conference as noted by (Wagner, Malisz, and Kopp, 2014).

The linguistic community witnesses also has a proliferation of multimodal resources for various purposes, for example, resources created only in the last year are Chat-talk Corpus (Yamazaki et al., 2020), AICO Multimodal Corpus (Jokinen, 2020), MULAI (Jansen et al., 2020), RDG-Map (Paetzel, Karkada, and Manuvinakurike, 2020), Chinese Whispers (Kontogiorgos, Sibirtseva, and Gustafson, 2020), MuSE (Jaiswal et al., 2020), Fakeddit (Nakamura, Levy, and Wang, 2019) etc. This strong interest is linked to the fact that accounting for verbal or textual information only does not suffice to provide a full picture of human communication.

However, not all languages are well represented in this kind of studies. According to LRE Map 5 there are currently 24 monolingual corpora for Italian, two of which concern spoken language, i.e. VoLIP (Alfano et al., 2014) and LUNA corpus (Dinarelli et al., 2009), and one multimodal, named ImagAct- ItalWorNet-Mapping (Bartolini et al., 2014); no entry includes an Italian corpus for the political domain.

Furthermore, researchers in Italian politics have mainly focused on political communication in the verbal modality, evaluating monological

discourse (Bolasco, de'Paratesi, and Giuliano, 2006; Cedroni, 2010; Longobardi, 2010; Catellani, Bertolotti, and Covelli, 2010; Bongelli, Riccioni, and Zuczkowski, 2010; Zurloni and Anolli, 2010; Sprugnoli et al., 2016) to study a politician's lexical, textual or rhetorical patterns. An exception is the work by (Salvati and Pettorino, 2010), that diachronically analyses some of the suprasegmental aspects of Berlusconi's speeches from 1994 to 2010. The corpus, however, is not available for further studies. Concerning political corpora developed specifically for conversation analysis, (Bigi et al., 2011) present a multimodal corpus of political debates at the French National Assembly, on May 4th, 2010 and introduce an annotation scheme for a political debate dataset which is mainly in the form of video and audio annotations. (Navarretta and Paggio, 2010) deal with the identification of interlocutors via speech and gestures in annotated televised political debates in British and American English.

Other papers have focused primarily on visual aspects (gaze, gestures, facial expressions) of communicative interaction during political talk shows or parliamentary speeches (D'Errico, Poggi, and Vincze, 2010).

Finally, the work presented in (Koutsombogera and Papageorgiou, 2010) analyzes a Greek multimodal corpus of 10 face-to-face television interviews focusing on non-verbal aspects to study the attempts of persuasion and interruption during political interviews. This work has some common ground with ours, sharing the application domain and the type of gesture annotation. However, the research objectives are quite different. The work of (Koutsombogera and Papageorgiou, 2010) primary focus is the study of the strategies for conversational dominance, thus presenting a specific annotation only on those traits. Instead, our work is broader in scope, including a different set of tags, adding a new semantic annotation layer, and integrating automatic linguistic features.

Chapter 3

PoliModalCorpus 2.0.

In the context of a political interview, the host, typically a journalist, acts as a representative of the audience. This means that, if a politician manages to convince or deal with the criticism that the host addresses, then her/his trustworthiness, reliability and credibility will be easily established. In this situation, a politician is judged not only based on one's arguments and rhetorical choices, but also on the attitude, self-confidence, and in general on an overall convincing behavior. For example, if a politician seems to be conversationally dominant and manages interruptions to a satisfactory degree, it is more likely that the host, and therefore the audience, will be convinced by the arguments put forward by the interviewee. For this reason, analyzing the combination of verbal and non-verbal elements in a political interview could be very interesting for scholars in political and communication science, and in general to study consensus mechanisms.

In this light, we present the first multimodal corpus of political interviews in Italian and analyze how the combination of verbal and non-verbal elements can shed new light into political agendas and politicians' attitude. By 'multimodal' or 'multimedia' we mean that the corpus is composed by audio-video recordings of interviews broadcast on TV with an orthographic transcription, which aims at the transposition of speech into the standard of the writing system, assuming as a conventional reference entity the graphic word. Such corpora are used especially for pragmatic

research purposes (e.g., in studies of proxemic correlates of spoken language or the bodily manifestation of emotions), in which the source sessions consist of movies that are transcribed and annotated. (Cresti and Panunzi, 2013). In our case, the transcriptions are annotated with information not only about the linguistic structure of the utterances but also about non-verbal expressions.

The corpus, which we call PoliModal, addresses the need to make up for the lack of Italian linguistic resources for political-institutional communication and is annotated in XML following the standard for the transcriptions of speech: TEI Guidelines for Electronic Text Encoding and Interchange.

In all transcripts, interviewers, interviewees and other guests' turns have been enriched with the manual annotation of non-lexical and semi-lexical aspects such as breaks, interruptions, false starts, overlaps, interjections, etc. Furthermore, additional linguistic traits related to language complexity, use of pronouns and persons' mentions have been automatically tagged, enabling an in-depth analysis of speakers' attitude and communication strategy.

In this chapter we present not only the corpus, but also an analysis that, combining verbal and non-verbal elements, shows how these traits contribute to making an interview more or less convincing. This work is an extension of (Trotta et al., 2019) in that we provide more details on the annotation scheme and guidelines, including inter-annotator agreement. Furthermore, we introduce additional analyses concerning the relation between non-verbal traits and related linguistic content, showing for example that overlapping or repeated terms can shape different conversational styles and rhetorical strategies in the interviews.

3.1 Description of the PoliModal Corpus

The PoliModal corpus includes the transcripts of 56 TV face-to-face interviews of 14 hours - taken from the Italian political talk show “In mezz’ora in più” broadcast from 24 September 2017 to 14 January 2018. The show follows a fixed format, with interviews conducted by a journalist, Lucia Annunziata, to a guest, typically a prominent figure in the political or cultural scene. A secondary guest may participate as well, usually a second journalist to comment on the debate. Each interview is done in the same limited time frame, 30 minutes, and no audience is present, so that applause and any other type of reactions are not included in the corpus.

The audio signal has been transcribed using a semi-supervised speech-to-text methodology (Google API + manual correction). All hesitations, repetitions and interruptions of the original interview have been included. The output has been further segmented into turns, and punctuation has been added, mainly to delimit sentence boundaries when they were not ambiguous.

It is important to note that, even if transcription seems to be an objective task, it involves a certain degree of interpretation. Indeed, the inclusion of the punctuation necessary to make the writing comprehensible, as well as the selection of non-verbal messages and non-verbal expressions (interjections, laughter, unfinished words, etc.) are interpretative choices aimed at revealing a sense¹. Therefore, in the case of ambiguous sentences, they have been identified manually, mainly looking at the context of the enunciation. According to (Ducrot, 1995), in fact, it is not possible to understand a communicative act without knowing the context in which it occurs. The context is therefore essential to choose one of the possible interpretations of ambiguous expressions.

¹As (Portelli, 1985) reminds us: “La punteggiatura serve sia a scandire il ritmo che a gerarchizzare sintatticamente il discorso; non sempre le due funzioni coincidono, per cui trascrivendo si è costretti spesso optare per l’una a danno dell’altra.”

In PoliModal, annotation has been done using XML as markup language and following the TEI standard for Speech Transcripts in terms of utterances². The linguistic resource has currently 100,870 tokens and includes interviews to politicians covering all the Italian political spectrum (from the extreme right movement Casa Pound to the liberal and progressive Partito Radicale). Beside politicians, also a small number of people with different backgrounds (students, academics, judges, economists, etc.) has been interviewed and is therefore included in the corpus.

3.1.1 Annotation Scheme

In this section we will describe in detail the creation of the corpus and the annotation scheme used³.

The first step we made is the orthographic transcription. Source files are in .mp4 format, our goal is to obtain a plain text (.txt) encoded in UTF-8 (Unicode Transformation Format, 8 bit) in order to guarantee interoperability and prepare the text for further automatic analysis. Although there are many tools that can be used for transcription (e.g. PRAAT, OH Portal, OTranscribe, Transcribe etc.) we propose a semi-automatic speech-to-text methodology using Web Using API⁴ and manual correction in order to correct any errors caused by automatic transcription and insert punctuation.

At the end of the human review process, the text in .txt format and coded in UTF-8 will be ready to be annotated. It will then appear as follows, with a turn per line:

The example presented above as well as all the other examples reported in these guidelines are taken from the corpus.

²P5: Guidelines for Electronic Text Encoding and Interchange See more <https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html>TSSAP

³Guidelines are described in detail on our repository

⁴See more detail at <https://wicg.github.io/speech-api/>

```
Lucia Annunziata: E buongiorno Matteo Renzi, Segretario del Partito Democratico.  
Matteo Renzi: Buongiorno!  
Lucia Annunziata: Bentornato, è quasi un anno che lei non era qui in questo studio e noi  
ci prenderemo oggi un pò di tempo per tentare di riannodare anche un pò di fili di un  
discorso che per un pò di mesi non abbiamo fatto direttamente con lei. È una settimana  
molto importante non lei è impegnato in un lunghissimo viaggio di 107 tappe. Ne ha fatte  
solo 21, un tour de force.  
Matteo Renzi: Beh solo 21. La prima, la prima settimana, 21 tappe in treno...
```

FIGURE 3.1: Raw text

The annotation scheme wants to keep track of so-called “speech constants” (Voghera, 2001) (e.g. dialogical organization in turns, use of repetition, use of speech signals, etc.). Indeed, spoken language deviates systematically and regularly from the written form, and researchers (Biber, 1995; Miller, Miller, Weinert, et al., 1998) have observed that there are constants of speech that make two spoken texts in a natural and spontaneous context similar, even if belonging to different diastratic or diaphasic levels. Thanks to these constants, spoken texts tend to resemble each other more than a spoken and a written text belonging to the same diaphasic and diastratic level. This level of annotation is largely inspired by the TEI standard for spontaneous speech. However, the tags that we will present are not the same as those present in the standard. In some cases, functional changes have been made in compliance with the research objectives of the PoliModal corpus.

Metadata annotation

As recommended by TEI where a computer file is derived from a spoken text rather than written one, it will usually be desirable to record additional information about the recording or broadcast which constitutes its source. So, in the metadata we insert useful information for a quick identification of transcriptions, for example the duration (shortened to *dur*) of the registration taken into account; the equipment that provides technical

details of the equipment and media used for an audio or video recording used as the source for a spoken text; the title of the format of television program from which the interviews are taken; the main roles of the interview are indicated (interviewer and interviewee), followed by the name and surname of corresponding role; date of airing, etc. Below we show an extract from the interview to Matteo Renzi aired on 22 of October 2017.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE trascrptions SYSTEM "unicum.dtd">
<trascrptions>
  <teiHeader>
    <recording type="video" dur="01:01:46">
      <equipment>
        <p>Recorded from http://www.raiplay.it/video/2017/10/12-h-in-piu-35b932c7-9a17-4fba-9d85-2f71d91f5806.html</p>
      </equipment>
    </recording>
    <broadcast>
      <bibl>
        <title>political talk-show</title>
        <author>Rai 3</author>
        <respStmt>
          <resp>interviewer</resp>
          <name>Lucia Annunziata</name>
        </respStmt>
        <respStmt>
          <resp>interviewee</resp>
          <name>Matteo Renzi</name>
        </respStmt>
        <series>
          <title>In mezz'ora in piu</title>
        </series>
        <date when="2017-10-22">22 October 2017</date>
      </bibl>
    </broadcast>
  </teiHeader>
```

FIGURE 3.2: Metadata annotation

Utterances and speaking turns

As explained by TEI: “Most researchers agree that the utterances or turns of individual speakers form an important structural component in most kinds of speech, but these are rarely as well-behaved (in the structural

sense) as paragraphs or other analogous units in written texts: speakers frequently interrupt each other, use gestures as well as words, leave remarks unfinished and so on. Speech itself, though it may be represented as words, frequently contains items such as vocalized pauses which, although only semi-lexical, have immense importance in the analysis of spoken text. Even non-vocal elements such as gestures may be regarded as forming a component of spoken text for some analytic purposes”⁵.

Each transcription consists of alternating turns between the sender and receiver in the case of dialogic form. Therefore, annotators are asked to segment the document marking this kind of components. See for example the following excerpt:

```
<u who="Lucia Annunziata" role="host" gender="f">E qual è il profilo del nuovo  
Governatore?</u>  
<u who="Matteo Renzi" role="PD party secretary" gender="m">Vorrei che chiunque fosse  
scelto, fosse il o la migliore persona possibile. Il candidato o la migliore candidata  
possibile.</u>
```

FIGURE 3.3: Annotation of utterances

So *u* (utterance) contains a stretch of speech usually preceded and followed by silence or by a change of speaker and include the attributes: name and surname of who holds the turn; role that indicates the work of the person who holds the turn; gender.

Pausing

Speakers differ very much in their rhythm and in particular in the amount of time left between words or utterances. Several studies have converged on the conclusion that we alternate between planning speech and implementing our plans. Indeed, as shown in (Henderson, Goldman-Eisler, and

⁵P5: Guidelines for Electronic Text Encoding and Interchange. See more in paragraph 8.1 “General Considerations and Overview”: <https://www.tei-c.org/release/doc/tei-p5-doc/it/html/TS.html>

Skarbek, 1966), participants to interviews typically show a cycle of hesitation and fluency, although the ratio of speech to silence varies among speakers.

The pause tag is therefore used to mark when the speech has been paused, irrespective of the actual amount of silence. In our annotation this tag is used to mark a pause within utterances. Below we present an example of the breaks taken from the interview with Pier Carlo Padoan aired on 14 of January 2018.

```
<u who="Pier Carlo Padoan" role="Minister of Economy" gender="m">Siamo usciti da una
crisi, lo ricordava lei prima, che in<pause/>tre anni di recessione ha portato via quasi
10 punti di Pil.</u>
```

FIGURE 3.4: Pausing annotation

A pause marked within an utterance applies to the speaker of that utterance. The attribute type may be used to categorize the pause, for example as short, medium, or long; alternatively, the attribute dur (i.e. duration) may be used to indicate its length more exactly.

Vocal

A typical aspect of spoken language is the use of semi-lexical and non-lexical expressions. Lexical expressions consist mainly of interjections (lexical category that conveys the meaning of an entire sentence, so it alone constitutes a complete linguistic act demonstrated by the fact that it is paraphrasable). Instead non-lexical expressions consist of phenomena such as coughing, exhaling, sniffing. The presence of non-transcribed semi-lexical or non-lexical phenomena either between or within utterances is foreseen also by the TEI standard. It can be marked using the following tags: a) vocal marks any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc.; b) kinesic marks any communicative phenomenon, not necessarily vocalized,

for example a gesture, frown, etc.; c) incident marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication. In this first phase of annotation, the only phenomena we focused on is “vocal”, that – in the vast majority of cases – marks an interjection.

```
<u who="Matteo Renzi" role="PD party secretary" gender="m"><vocal type="semi-lexical"
desc="mm"/>Non so se sono stato il golden boy o l'antisistema. Io so soltanto che quando
vedo ed esco da un incontro con i terremotati di Arquata, o quando vedo ed esco da
un'azienda in crisi a<vocal type="semi-lexical" desc="ehm"/>sul treno incontriamo i
ragazzi<del type="repetition">della della</del>Perugina o<vocal type="semi-lexical"
desc="ehm"/>a Civita Castellana, il settore delle ceramiche dove in parte è forte quello
delle ceramiche sanitarie, e in parte è stato cancellato dall'avvento della Cina, quello
delle stoviglie. Quando io incontro queste persone, nessuno mi domanda del governatore
della Banca d'Italia ma tutti mi domandano come si fa ad avere mutui diversi.</u>
```

FIGURE 3.5: Vocal annotation

As can be observed from this excerpt, the vocal tag enables two types of attributes: type, which admits semi-lexical or non-lexical values, and desc, which may be used to supply a conventional representation for the phenomenon (non-lexical e.g. burp, click, cough; semi-lexical e.g. ah, ehm).

These traits have been associated with the fact that linguistic planning is very cognitively demanding, and it is difficult to plan an entire utterance at once (Lindsay, 1975). Therefore, hesitation pauses, and similar vocal phenomena may be useful to perform a careful lexical retrieval, since past studies (Levelt, 1983) found that pauses occurred more often before low-frequency words than before high frequency ones.

False starts, repetitions and truncated words

Phenomena of speech management include disfluencies such as filled and unfilled pauses, interrupted or repeated words, corrections, and reformulations as well as interactional devices asking for or providing feedback. These phenomena are marked as editorially deleted i.e. del in the annotation.

Although spoken texts are the product of a physically continuous process, their structure shows a strong discontinuity: false starts, interruptions, project changes are common to all spontaneous speech texts. In particular by false start we mean the abandonment by the speaker of a word or a sequence of words already produced, with or without repetition of the previously used linguistic material (Cresti and Panunzi, 2013). The false starts are therefore noted as follows:

```
<u who="Matteo Renzi" role="PD party secretary" gender="m">Il migliore<del
type="falsestart"/>Questa è una valutazione che deve fare il Presidente del Consiglio.
</u>
```

FIGURE 3.6: False Start annotation

As noted by (Voghera, 2001, p. 7): *“Si è da più parti notato (Simone, 1990; Bazzanella, 1992; Tannen, 1989; Voghera, 1992) che nel parlato spontaneo vi è un’alta percentuale di ripetizioni. [...] Esistono infatti vari tipi di ripetizione con funzioni diverse, che possiamo ricondurre a due macrocategorie (Voghera, 1992): ripetizione di enunciati altrui per dare coerenza e coesione al discorso; autoripetizione di tipo automatico come meccanismo di controllo della programmazione del discorso. Tanto il primo quanto il secondo tipo di ripetizione sono funzionali al controllo della progettazione testuale in fieri del parlato”.*

(en. Many works note that (Simone, 1990; Bazzanella, 1992; Tannen, 1989; Voghera, 1992) that in spontaneous speech there is a high percentage of repetition. [...] In fact, there are various types of repetition with different functions, which can be traced back to two macro-categories (Voghera, 1992): repetition of other people’s statements in order to give coherence and cohesion to speech; repetition of him/herself as a mechanism for controlling the programming of speech. Both the first and the second type of repetition are functional to the control of the verbal design in progress of speech.)

The type of repetition does not affect the tag used, which will be unique, as follows:

```
<u who="Lucia Annunziata" role="host" gender="f">Dunque<del type="repetition">lei dice,  
lei dice</del>che non è sbagliato.</u>
```

FIGURE 3.7: Annotation of repetition

Spontaneous dialogical texts present a frequent and somewhat ‘compulsory’ use of deictic elements (Givón, 1995). Some forms of ellipses can also be traced back to deictic or indessical phenomena (Berretta, 1994). The same need for indessicality can also be traced back to cases of reduction and truncation. Truncations - which also fall into the category of editorially deleted - are annotated as follows:

```
<u who="Lucia Annunziata" role="host" gender="f">E per esempio sulla legge elettorale di  
reintrodurre preferenze sulle liste bloccate prof<del type="truncation"/>proporzionali o  
aumentare i nominati.</u>
```

FIGURE 3.8: Truncation annotation

Overlap

We include this phenomenon in our annotation since several past studies (Simone, 1990; Bazzanella, 1992; Tannen, 1989) highlighted their importance in spontaneous speech, mentioning in particular the role of repetitions in controlling the in-progress textual design of speech (Voghera, 2001). As (Voghera, 2001) points out again, the conditions of construction and reception of texts mean that speech needs a lot of redundancy because it is more exposed to noise than writing. A source of noise is the alternation of turns, which can lead to overlapping of the participants in the communication and, therefore, partial or total loss of information. So this phenomenon is present when the speaker conveys (in a verbal or non-verbal manner) that he/she is about to finish his/her turn and the co-locutor starts speaking so that there is a slight overlap of utterances. Overlays will be annotated as follows:

```

<u who="Matteo Renzi" role="PD party secretary" gender="m">E con il debito che cala, è
un'operazione che...</u>
<u trans="overlap" who="#Lucia Annunziata" role="host" gender="f">Che si può fare.</u>
<u who="Matteo Renzi" role="PD party secretary" gender="m">Che i più grandi gruppi
internazionali sottoscrivono eh.</u>

```

FIGURE 3.9: Overlap annotation

In the example above, the journalist Lucia Annunziata breaks Matteo Renzi’s turn to speak, completing the sentence through an overlap. The simplest way of representing this overlap is to use the `trans` attribute that is provided as a means of characterizing the transition from one utterance to the next. The tag is completed by entering the name and surname of the person who overlaps preceded by `#`, then adding role and gender below.

Overlaps can be competitive, when the overlapper disrupts the speech and can be perceived as intrusive by dominating the conversation, and cooperative, when the goal of the overlapper is to maintain the flow of the turns and add to the conversation with further comments (Truong, 2013).

Inter-annotator agreement

The annotation task addressed so far falls – from a qualitative point of view – in the first of the general types identified by (Mathet, Widlöcher, and Métivier, 2015), in which the subjective interpretation is limited. Indeed, it deals with the “identification of units” (Krippendorff, 2018), in which the annotator, given a written or spoken text, must identify the position and boundary of linguistic elements (e.g. identification of prosodic or gestural units, topic segmentation).

In order to evaluate the reliability of our annotation scheme, we compute inter-annotator agreement by performing a double annotation of verbal and non-verbal traits of the first ten minutes of Renzi’s, Di Maio’s

and Salvini’s interview. Both annotators were expert linguists. Macro-averaged F1 computed on exact matches amounts to 0.82, which corresponds to a good agreement, given that by exact match we consider the correct choice of the trait, the position of the tag and the exact extension of the marked string, if any. This result confirms the reliability of the task and the corresponding annotation guidelines.

3.2 Corpus Analysis

In this section, we analyze several linguistic dimensions that can be either automatically extracted or derived from the corpus annotation, and that can contribute to better understand typical traits of political communication. Overall statistics for the traits annotated for each politician in the corpus are displayed in Table 3.1.

<i>Guest</i>	<i>Turn</i>	<i>Repet.</i>	<i>FalseSt.</i>	<i>Trunc.</i>	<i>Overlap</i>	<i>Pause</i>	<i>Non-lex.</i>	<i>Semi-lex.</i>
Alessandro Di Battista	203	24	14	34	76	19	9	66
Carlo Calenda	137	10	13	1	48	37	1	34
Matteo Renzi	187	40	19	69	25	0	3	16
Walter Veltroni	55	16	12	10	11	0	2	8
Simone Di Stefano	91	20	5	15	23	0	0	4
Pierluigi Bersani	92	30	0	20	15	1	14	24
Angelino Alfano	100	17	3	3	31	9	2	22
Giulio Tremonti	56	8	0	0	14	9	2	6
Matteo Orfini	67	10	0	0	21	1	2	8
Luigi Di Maio	74	14	0	14	32	0	4	11
Matteo Salvini_1	57	13	0	11	19	3	2	14
Matteo Salvini_2	86	19	3	3	30	13	7	19
Pier Carlo Padoan	67	5	1	7	13	8	13	21

TABLE 3.1: Corpus statistics related to the 13 interviews included in our study

3.2.1 Statistics of Non-Verbal Traits

Recent studies in political communication have showed that non-verbal cues have different impact depending on the political orientation of the source and of the receiver (Laustsen and Petersen, 2016). Along this line, we analyse whether the political orientation significantly influenced the production of non-verbal items and the communication strategies used by the interviewees. We therefore group the politicians in our corpus into political parties, and then analyze those that are represented by least 3 politicians: Forza Italia, a conservative center-right political party (3 interviews), Lega Nord, a right-wing political party often targeting immigrants (5 interviews), Movimento 5 Stelle, a populist citizens' movement (3 interviews) and Partito Democratico, a moderate centre-left political party (9 interviews). An overview of the distribution of non-verbal traits in the PoliModal corpus for each party is reported in 3.10. Although the graph shows some differences in the frequency of occurrences, they are not statistically significant, also because of the relatively small number of interviews considered in the study. Also, the standard deviation for the averages tends to be high, showing high differences among interviewees of the same party. For example, politicians of Lega Nord make on average more pauses, but the range goes from 0.286 per turn (Roberto Maroni) to 0 (Luca Zaia). Similarly, non-lexical and semi-lexical expressions, marked as vocal, are on average more frequent for PD politicians, but range from 1.25 per turn (Enrico Letta) to 0.10 (Matteo Renzi). These results show that differences pertain more to single persons and conversational style than to political orientation. An exception is given by overlaps, for which the three politicians of M5Stelle (Alessandro Di Battista, Luigi Di Maio, Giancarlo Cancelleri) all show a frequency above average, suggesting that it may be connected with the communication strategy of the members of Movimento.

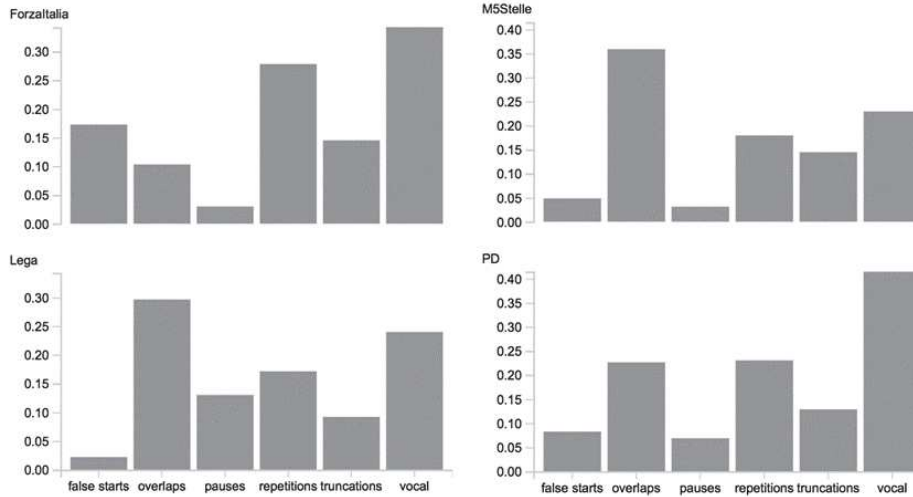


FIGURE 3.10: Distribution of traits per political party (avg. number of occurrences per turn)

3.2.2 Political Orientation and Language Use

A second analysis we carry out is related to existing works about the use of linguistic features related to political orientation. In particular, a recent study by (Schoonvelde et al., 2019) has analyzed more than 380,000 speeches from five different Parliaments and has proven that ideologically conservative politicians use a less complex language than liberal ones. Since these findings were not tested on Italian political documents, we carry out a comparison using the collected transcripts. In order to analyze the complexity of the language used by each politician we computed the type-token ratio and the average lexical density, i.e. the number of content words divided by the total number of tokens. We do not take into account the Gulpease index (Lucisano and Piemontese, 1988) which is the de-facto standard metric of readability in Italian, because it was meant for written documents and heavily relies on sentence length, a boundary that is not always present in transcripts.

Considering the average type-token ratio and conceptual density per

political party, there are almost no variations among the parties, with small standard deviations. Indeed, conceptual density ranges between 0.58 (avg. PD and Movimento 5 Stelle) and 0.59 (avg. Forza Italia and Lega), while type-token ratio ranges from a minimum of 0.74 (PD) and a maximum 0.78 (Movimento 5 Stelle). This comparison suggests that in our case the hypothesis by (Schoonvelde et al., 2019) is not confirmed, with the three highest type-token ratio values belonging to politicians from three different parties: Forza Italia (Mariastella Gelmini, 0.87 ttr), Lega Nord (Matteo Salvini, 0.82) and PD (Michele Emiliano, 0.82).

A second hypothesis we want to test is the one introduced in the work by (Cichocka et al., 2016), where the authors show that Republican presidents used a higher proportion of nouns than Democratic presidents, while there were no reliable differences in the use of verbs or adjectives. The authors suggest that, compared to liberals, conservative politicians are more inclined to use parts of speech that stress clarity and predictability (such as nouns) and reduce uncertainty and ambiguity (such as verbs or adjectives).

We therefore compute the average number of nouns, adjectives and verbs per political party and compare them. Similar to the previous analysis, averages are all in the same range and there is no statistically significant difference among parties. However, some of the results are in line with Cichocka et al.'s study, with PD showing a slightly lower number of nouns on average (and Valeria Fedeli being the politician with the lowest noun ratio, 0.16). Also, Matteo Salvini and Luigi di Maio are the politicians with the highest use of nouns, 0.22 per token on average. A further evidence in favour of these results are the statistics obtained on the use of content words, in particular on the percentage of nouns, verbs, adverbs and adjectives, reported in Fig. 11. We consider the five politicians with the highest number of turns in the corpus

: Alessandro Di Battista (Movimento 5 Stelle), Carlo Calenda (PD),

Matteo Renzi (PD), Angelino Alfano (Popolo della Libertà), Matteo Salvini (Lega). The figure confirms that Matteo Salvini is the politician using the most number of nouns on average compared with other interviewees, in line with the findings by (Cichocka et al., 2016). Carlo Calenda, instead, is the politician who on average uses most verbs and adverbs, conveying more uncertainty and ambiguity than all the other politicians including Matteo Renzi.

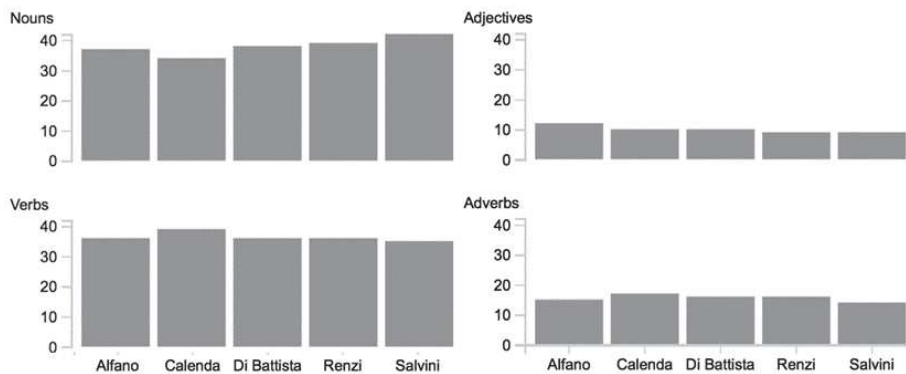


FIGURE 3.11: Use of nouns, verbs, adjectives and adverbs for each politician (% over all content words)

Finally, we analyze how the speakers in our corpus make use of the personal pronouns “io” (I) and “noi” (we), along the line of similar studies carried out on political discourse in English (Chilton, 2004). Since Italian is a pro-drop language, we include in the analysis also information extracted from the verb morphology, in particular the use of first person singular and plural. We focus on the five politicians included in 3.11 plus Luigi di Maio, who at the time was the head of Movimento 5 Stelle. The extraction of pronouns and of the verb morphology is carried out by processing the corpus with TINT (Palmero Aprosio and Moretti, 2018), a widely used suite for Italian text processing.

Pronouns are more relevant because they mark the person and they are less prone to analysis mistakes, although with TINT some ambiguous cases may be wrongly marked as first person singular). We observe that Angelino Alfano shows the highest incidence in the use of “io” (I) (0.60 per turn), followed by Matteo Renzi and Carlo Calenda (both 0.51 per turn). On the contrary, the use of “noi” (we) has the highest incidence in Di Maio’s interview (0.25 per turn), who uses the plural form to stress the role of Movimento as a collective entity in which decisions and opinions are shared, as opposed to the ego-centric political discourse of Alfano and Renzi (0.10 and 0.13 respectively). Alessandro Di Battista, despite belonging to the Movimento, does not speak in the name of others, indeed he makes a limited use of “noi” (only 0.09 per turn). If we include in the analysis also inflected verbs, we observe that terms in the first-person singular are used most by Angelino Alfano, 5.5 times more frequently than expressions in the first person plural. Luigi di Maio, instead, uses the first person singular only 1.4 times more than the plural form, confirming the finding that this may be part of a communication strategy of Movimento. However, this is not shared by Di Battista, that, although belonging to the same party, seems to speak only in his name (singular form used 4.2 times more than plural one).

The fact that the three studies considered do not find a clear confirmation in our corpus, where the differences among the parties are rather blurred, may have three possible explanations:

- this corpus may be too small to test the above hypotheses. Its expansion is indeed already in progress;
- the hypotheses do not actually hold in our case, i.e. in the Italian political scene it is not true that liberals use more complex language and tend to use less nouns than conservatives;

- the four parties considered cannot be straightforwardly divided into liberals and conservatives, and there are different positions inside the same party.

3.2.3 Relation Between Verbal and Non-Verbal Traits

A third analysis is aimed at studying the correlation between non-verbal traits and language complexity. We, therefore, focus on interviews that have a minimum length of 50 turns. The list of politicians and the corresponding count of annotated traits is reported in Table 3.1.

Again, for complexity, we consider type-token ratio and conceptual density. We perform an analysis of the correlation between language complexity and the six non-verbal traits manually annotated in the interviews, normalized by the number of turns uttered by each politician. While type-token ratio (TTR) does not correlate with any of the manual traits, we found that lexical density shows a moderate negative correlation with repetitions ($n=13$, $r = -0.51$), truncations ($r = -0.46$) and non-lexical and semi-lexical expressions ($r = -0.43$). On the contrary, it has a moderate positive correlation with the average number of pauses ($r = 0.49$). This result suggests that, among the manual traits, pauses are used as a linguistic device and are an indicator of a good control of the conversation. Therefore, they are more often used by politicians showing a high lexical density, i.e. the ability to convey concepts in a concise way, which is crucial especially during TV interviews. The other manually annotated traits, instead, seem to be more frequent in speeches that are less organized, for which the management of the discourse is less efficient.

Among the politicians considered in this study, Carlo Calenda makes on average the highest number of pauses (0.27 per turn on average, with a lexical density of 0.57), followed by Giulio Tremonti (0.16 pauses per turn, 0.58 lexical density).

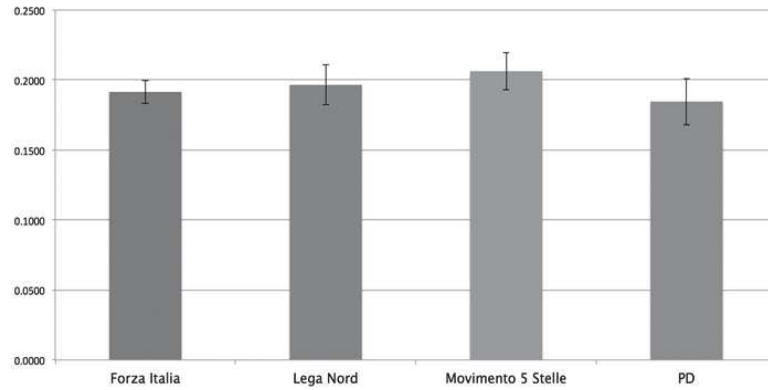


FIGURE 3.12: Avg. nouns per political party

Finally, we perform a qualitative analysis of the content of turns correlated with non-verbal traits. In particular, we analyze the linguistic content of strings being part or immediately preceding repetitions, truncations and overlaps, to check whether there are specific patterns associated with the different speakers. Concerning *Repetitions*, we observe that most speakers repeat function words such as articles, connectives, prepositions, adverbs, using repetitions as a device to take time, while conveying an overall attitude of uncertainty. There are however few exceptions: Matteo Renzi tends to repeat expressions highlighting his actions and opinions (e.g. “io dico, io dico” (I say, I say); “io ho detto, io ho detto” (I said, I said); “rivendico, rivendico” (I claim, I claim); “guardo, guardo” (I look at, I look at)). Matteo Salvini seems to be the politician that most uses repetitions as an emphatic device, to stress key items in his political agenda. Indeed, he repeats not just single words or expressions but also full sentences, see for example “La legge Fornero non va cambiata in 5 anni, la legge Fornero non va cambiata in cinque anni” (En. “Fornero law should not be changed in five years, Fornero law should not be changed in five years”); “Sicuramente inferiore al 20%, sicuramente inferiore al 20%”; “ti chiedo il 15%, ti chiedo il 15%” (En. Surely less than 20%, surely less than 20%”; “I ask you for 15%, I ask you for 15%.”). His use of repetitions is very

effective in conveying certainty and emphasis, as opposed to repetitions expressing indecision by other interviewees.

As regards *false starts*, they are used by the interviewees to correct the wording of their utterances or rephrase expressions. In some cases, they highlight an (unconscious) rhetorical strategy of the speaker. For example, Carlo Calenda starts a turn by saying “Questo paese è nato sull’industria e sugli op-” but then corrects his wording by saying “sul lavoro e sulle imprese”, probably because “op[erai]” is too politically loaded and he implements a form of self-censorship. Other interesting examples are by Matteo Renzi, who restarts his utterance with the goal to involve the interviewer and move the discourse focus from himself to someone else, as shown below:

- *io uscivo costantemente sulla storia dei* *avevo i giornalisti che giustamente facevano il loro lavoro e ad ogni fermata mi chiedevano di Banca d’Italia* (En. I was constantly going out on the I had journalists who were rightly doing their job and at every stop they asked me about the Bank of Italy)
- *Non ne avevo* *Lei non ci crederà ma non avevo dubbi.* (En. I had none You won’t believe it but I had no doubts)
- *Insomma, non* *Lei non mi troverà mai fare una polemica con i giornali* (En. In short, I did not You will never find me making a controversy with newspapers)

Concerning *overlaps*, several differences can be observed comparing the interviews. In this case, also the interviewer’s attitude and opinion on the politician affects the conversation structure and the frequency of overlaps. Following the overlap analysis proposed by (Schegloff, 2000), we observe that overlaps can be seen as a collaborative oriented simultaneous talk in

Calenda's interview, who shows to be in control of the conversation and uses overlaps mostly to provide additional information in a constructive manner. Salvini shows a different attitude and uses overlaps mainly as a fight-for-floor device, with the goal to interrupt and correct the interviewer's statements. This is confirmed by the fact that his overlaps start mainly with negations and adversative conjunctions ("Ma") (But). With Di Battista overlaps are mainly competitive, leading to long chains of overlapping turns, where also the interviewer plays an active role in competing for the turn space. With Di Maio overlaps are often a response to negative remarks by the interviewer, against which the interviewee attempts a defense. Also in this case, overlaps start often with negations and adversatives and address directly the interviewer ("Sa benissimo dottoressa Annunziata", "Lei è bravissima a fare domande provocatorie", "No le dico", "Atteniamoci, guardi le dico, le chiedo questa cosa") (En. "You know very well Dr. Annunziata," "You are great at asking provocative questions," "No I tell you," "Let's stick to it, look I tell you, I ask you this thing.")

3.3 Multimodal Annotation

In addition to the level of annotation already present that is useful for studying the dialogical style of the interviewees, a further one has been added in order to enrich it with an additional mode and therefore a new level of meaning, expressed through facial displays, hand gesture and body posture. Adding this kind of information is very time-consuming, since it requires that the annotator watches the video interviews and marks traits derived from the video, while aligning them to the underlying text which was already transcribed.

In a first phase, the novel annotation was extended only to a subset of 3 interviews with three politicians belonging to different political parties (Matteo Renzi, from the center-left party Partito Democratico, Matteo

Salvini, from the right-wing party Lega, and Luigi di Maio, from the populist party Movimento Cinque Stelle) in order also to verify its reliability. When the interviews took place, they were candidates for the presidency of the Council of Ministers.⁶

Being therefore competitors on the Italian political scene, they had to establish an image for themselves as competent personalities, a goal which is considered equally important to the topic under discussion (Koutsombogera and Papageorgiou, 2010). At the same time, they had to respond to the interviewers' challenges and comments presenting their arguments and opinions in a persuasive way.

In the paper by (Allwood, 2008), the author highlights that synchronization of information in different modalities is a crucial issue in assembling a multimodal corpus. Therefore the authors suggest to adopt the general principle of spatio-temporal contiguity. This means that a text occurs at the same point in time as the event it describes or represents. When temporal contiguity concerns the relation between transcribed speech (or gesture) and recorded speech (or gesture), it is often referred to as "synchronized alignment" of recording and transcription. What synchronization means is that for every part of the transcription (given a particular granularity), it is possible to hear and view the part of the interaction it is based on and that for every part of the interaction, it is possible to see the transcription of that part. The form of connection between the transcriptions and the material in the recordings can vary from just being a pairing of a transcription and video or audio recording, where both recording and transcription exist but they have not yet been synchronized, to being a

⁶The Italian political elections referred to in the paper were held on Sunday, March 4, 2018. They followed the dissolution of the Chambers, which took place by decree of the President of the Republic Sergio Mattarella on December 28, 2017, a short time before the natural expiry of the 17th legislature, scheduled for March 14, 2018. The results saw the centre-right establish itself as the most voted coalition, with about 37 percent of the preferences, while the single most voted list, the Movimento 5 Stelle, collected more than 32 percent of the votes.

<i>Behaviour attribute</i>	<i>Behaviour value</i>
General face	Smile, Laugh, Scowl , Other
Eyebrow movement	Frown, Raise , Other
Eye movement	Extra-Open, Close-Both , Close-One, Close-Repeated , Other
Gaze direction	Towards-Interlocutor, Up, Down, Sideways , Other
Mouth openness	Open mouth , Closed mouth
Lip position	Corners up, Corners down, Protruded, Retracted
Head movement	Down, Down-Repeated , BackUp, BackUpRepeated, Other
Handedness	Both hands, Single hands
Hand movement trajectory	Up, Down, Sideways, Complex , Other
Body posture	Towards-Interlocutor, Up, Down, Sideways , Other

TABLE 3.2: List of gestures, following the list described in (Allwood et al., 2007). The presence of bold means that the gesture has been found in our dataset.

complete temporal synchronization of recordings and transcription.

In our case, audio and video signals as well as the annotations have been temporally synchronized by hand. Although the most convenient solution for synchronization is to carry it out using a computer program already while making the recording (see for example the AMI project⁷, and the CHIL project⁸), we did it manually since the recording and transcription of the corpus were done before knowing what layers would be exactly annotated.

The video annotation was carried out using the ANVIL tool (Kipp, 2001) while the levels and labels used in the annotation scheme are mainly inspired by the MUMIN coding scheme notation (Allwood et al., 2007).

Table 3.2 summarizes the list of gestures, as described in (Allwood et al., 2007).

The annotation – made at the moment by a single expert annotator – follows the criterion highlighted by (Allwood et al., 2007), claiming that

⁷<http://www.amiproject.org>

⁸<http://chil.server.de/servlet/is/101/>

annotators are expected to select gestures⁹ to be annotated only if they have a communicative function. In other words, gestures are annotated if they are either intended as communicative by the communicator (displayed or signalled) (Allwood, 2001), or judged to have a noticeable effect on the recipient. For example, mechanical recurrent blinking due to dryness of the eye might not be annotated because it does not seem in a given context to have a communicative function.

Regarding the annotation guidelines - as specified in (Allwood et al., 2007) - the attributes concerning the shape or dynamics of the observed phenomena are not fine-grained, because they only seek to capture features that are significant with respect to the functional level of the annotation. Once a gesture has been selected by an annotator because of its communicative role, it is annotated with functional values, as well as features that describe its behavioural shape and dynamics: this is what we call the modality-specific annotation level. An additional, multimodal annotation level concerns the relation that the gesture has either with other gestures or with the speech modality. The scheme provides a number of simple categories for the representation of multimodal relations. However, it does not include tags for the specific annotation of verbal expressions since its focus is on the study of gestures, which is why we have integrated them in order to study - in the future - the relationship between verbal and non-verbal expressions.

Following this principle, we do not annotate all gestures, focusing on what follows:

(a) **Facial displays:** they refer to timed changes in eyebrow position, expressions of the mouth, movement of the head and of the eyes (Cassell et al., 2000). The coding scheme includes features describing gestures and movements of the various parts of the face, with values that are either

⁹(Duncan, 2004) defines a gesture as a movement that is always characterised by a stroke, and may also go through a preparation and a retraction phase. Each stroke corresponds in MUMIN to an independent gesture.

semantic categories such as Smile or Scowl or direction indications such as Up or Down.

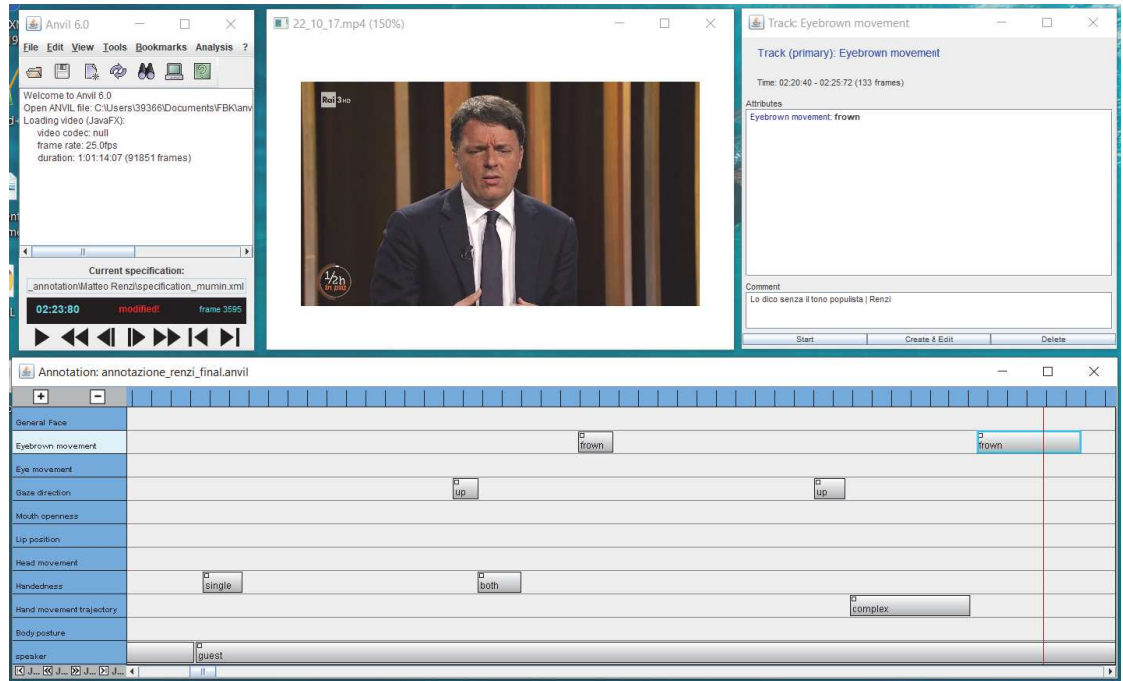


FIGURE 3.13: Example of facial display: frowning.

As an example, we report in 3.13 the annotation of the interview to Matteo Renzi. The leader of Partito Democratico frowns when discussing the defeat of his proposal in the constitutional referendum, at minute 00 : 02 : 23 : 80. This gesture, which - according to (Poggi, 2005) - can take on four main meanings (surprise, emphasis, contrasting, perplexity/doubt), takes here a contrasting meaning, because it occurs when the politician expresses his disagreement with what the interviewer just said about the referendum. Renzi's words uttered when making this facial expression are:

“Però, giusto per non perdere l'abitudine, non è che sia d'accordissimo sulla lettura che lei dà, nel senso che il referendum l'ho perso io.”

En: *“But - just so as not to lose the habit - I don’t agree with your interpretation, that is, I lost the referendum.”*

(b) **Hand gesture:** we follow a simplification of the scheme from the McNeill Lab (Duncan, 2004). The features, 7 in total, concern Handedness and Trajectory, so that we distinguish between single-handed and double-handed gestures, and among a number of different simple trajectories analogous to what is done for gaze movement. The value Complex is intended to capture movements where several trajectories are combined.

In 3.14 we show an example annotation of hand movements, in particular the use of both hands. At minute 00 : 01 : 55 : 72 Matteo Renzi, still discussing the defeat at the referendum, uses both hands – which could assume a batonic value ¹⁰ in this circumstance – while uttering the following sentence:

“Io quei politici che tutte le volte danno la responsabilità, la colpa, si nascondono dietro gli alibi personalmente non li sopporto.”

En: *“Personally, I can’t tolerate those politicians who always blame and hide behind alibis.”*

(c) **Body posture:** this tag comprises trajectory indications for the movement of the trunk. The categories are mutually exclusive to facilitate the annotation work.

3.15 shows a third example – taken again from the interview to Matteo Renzi, at minute 00:00:41:12 – in which the position of the interviewee’s bust appears slightly sideways. In this case, the gesture occurs while the interviewee listens to a question and therefore outside of a sentence. This annotation is therefore temporally aligned with the transcribed turn of the journalist.

¹⁰According to (Allwood et al., 2007): *“baton gestures are those in which the hands move rhythmically from top to bottom to scan and emphasize the accented syllables in a sentence”*.

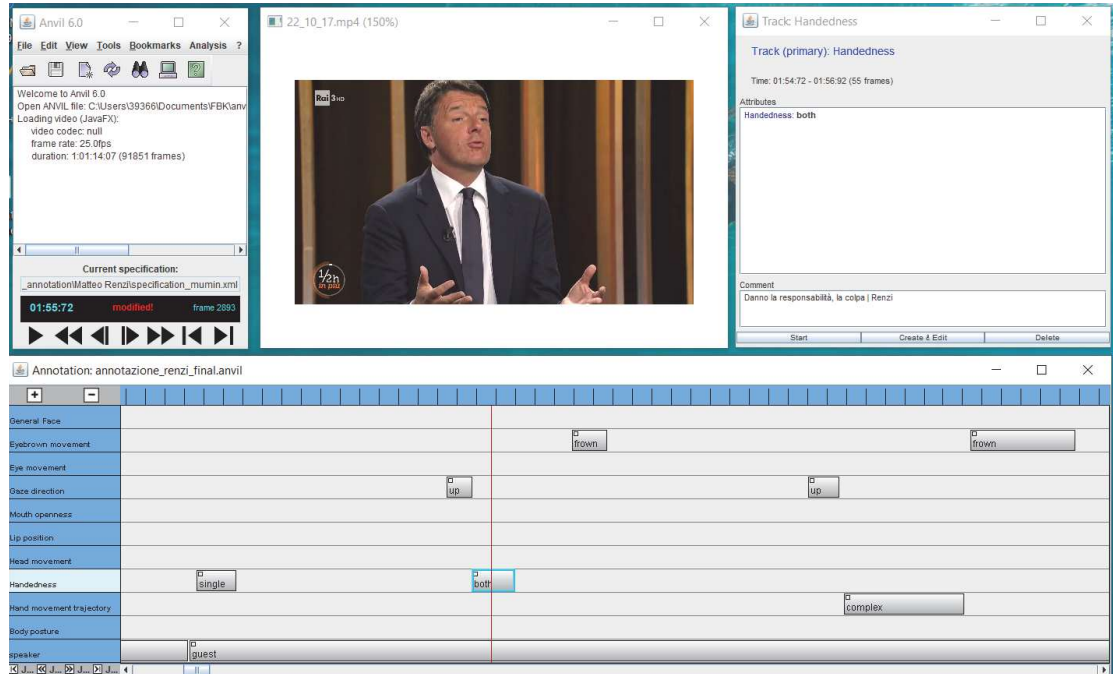


FIGURE 3.14: Example of hand gesture: double-handed.

3.4 Corpus Statistics and Discussion

Before presenting the quantitative results of the annotation extended to the entire corpus, we report below some qualitative-quantitative investigations performed on the interviews described above. This is also in order to demonstrate the innumerable potentialities offered by multimodal corpora. 3.3 shows the number of turns and the overall duration of the interview for each politician. The duration refers only to the interviewees' utterances, therefore excluding the time used by the journalist to make questions. The interviews to Luigi Di Maio and Matteo Salvini have a comparable duration both in terms of time and of turns. The interview to Matteo Renzi, instead, is longer (1 hour in total) but the turns are considerably shorter because he was being interrupted more frequently by the interviewer.

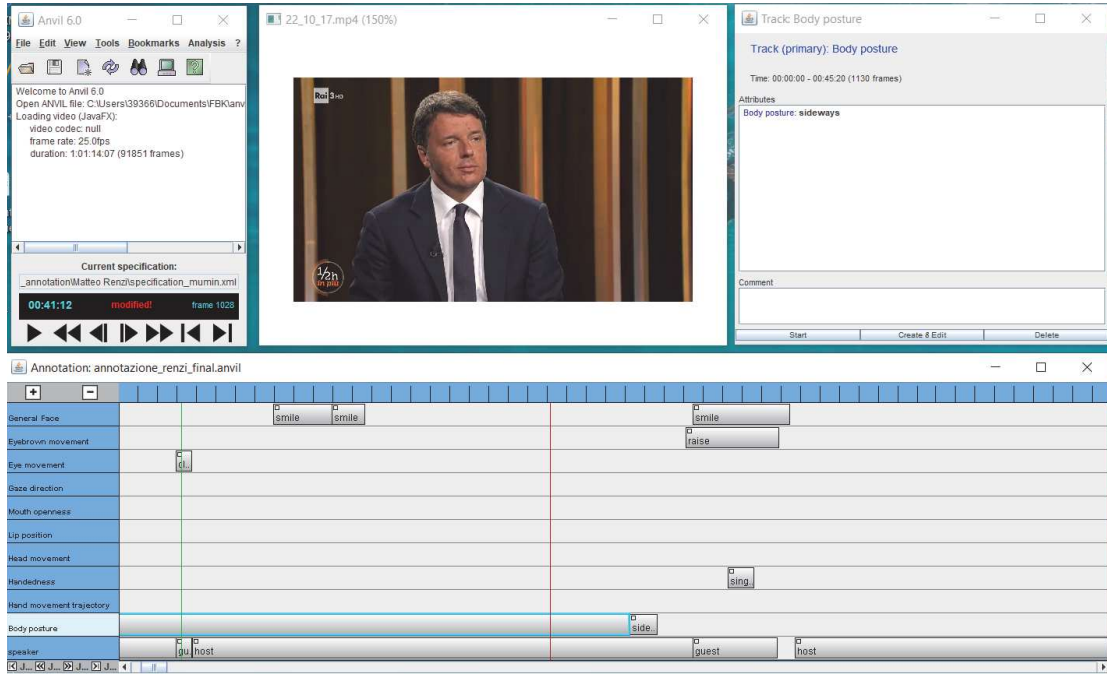


FIGURE 3.15: Example of body posture: sideways.

As regards the new annotation layer, we report in 3.4 the statistics for all annotated phenomena in the three interviews. Some traits that are present in the annotation scheme have not been reported because they have not been observed in any of the three interviews. For example, no occurrences of the extra-open and close-one eye movement types have been observed, nor the scowl among the facial expressions. Overall, Matteo Renzi shows the highest expressiveness through the use of gestures, facial displays and posture, with more than double occurrences compared to the opponents.

An interesting phenomenon is the movement of eyebrows, which has been extensively discussed also in the literature. In particular the frowning of the eyebrows, which as (Poggi, 2006) suggests indicates the rapprochement of the eyebrows, forming vertical wrinkles on the forehead,

<i>Politician</i>	<i>Turns</i>	<i>Duration (sec.)</i>
Matteo Renzi	149	2143.32
Luigi Di Maio	30	1113.92
Matteo Salvini	29	1070.28

TABLE 3.3: Corpus content: turns per speaker and total duration

may be used for a range of purposes, such as asking a question, communicating to an interlocutor that that s/he is not clear, expressing indirectly disagreement with the other party, looking at something very carefully, trying to remember something, asserting something with confidence, expressing concern or anger about something, giving a peremptory order.

In our specific case, the politician who shows the most frowning (30) is Matteo Renzi, and from the context we can argue that this signal is used by the former Prime Minister to show confidence in his assertions and exhibit attention to what is being said. The raising of the eyebrows – defined by (Purpura and Hillard, 2006) as “a signal of the gaze that is produced by lifting both eyebrows in a symmetrical manner” – may instead take on four main meanings: surprise, emphasis, adversity, perplexity/doubt. The semantic element shared by all these interpretations is the presence of new information, as a matter of unexpected knowledge. In surprise and in adverse meaning, a knowledge in entry is contrary to existing knowledge and therefore cannot be inferred based on the current state of things.

Overall, the different communication strategies adopted by the three politicians are evident in the corpus: Matteo Renzi’s gesture, facial displays and body posture express an extrovert attitude, but also an evident attempt to please the audience and to be convincing at all costs. This is confirmed also by the lexical and semi-lexical traits annotated in this interview that include a high number of repetitions and truncations (0.21 and 0.37 per turn on average, respectively) and no pauses, as if the interviewee could not organise well the discourse and was too much involved

	<i>Matteo Renzi</i>		<i>Luigi Di Maio</i>		<i>Matteo Salvini</i>	
	count	duration	count	duration	count	duration
Face						
laugh	9	51.2	7	40.56	1	4.04
smile	32	163.96	13	185.20	7	36.20
scowl	2	43.96	0	-	0	-
Eyebrow movement						
frown	30	120.8	4	53.20	0	-
raise	20	126.08	0	-	0	-
Eye movement						
close-both	4	7.76	0	-	0	-
close-repeated	10	56.6	2	61.56	0	-
Gaze direction						
up	3	3.36	0	-	0	-
sideways	2	7.52	0	-	0	-
towards-interlocutor	4	47.92	0	-	0	-
down	6	11.48	0	-	0	-
Mouth openness						
open	2	2.96	0	-	0	-
Head movement						
down-repeated	3	6.56	0	-	1	3.12
Handedness						
single	4	9.20	4	109.20	1	0.72
both	17	83.32	4	82.92	0	-
Hand movement trajectory						
complex	42	672.52	8	226.32	20	989.72
up	5	13.80	0	-	4	22.96
sideways	13	107.56	5	103.28	2	4.12
down	3	11.56	1	4.52	0	-
Body posture						
sideways	2	46.6	0	-	0	-
down	1	0,4	0	-	0	-

TABLE 3.4: Statistics on annotated information comparing number of occurrences and average duration in milliseconds.

in trying to convince the audience.

On the contrary, Luigi di Maio shows only 0.19 repetitions and 0.19 truncations per turn on average, while gaze, head and eye movements are almost not present. The only traits that are more present in his speech than in the others' are facial displays to convey a positive attitude through smiles and laughs. As for other lexical features, he makes a remarkably higher use of overlaps, 0.43 per turn (vs. 0.13 for Renzi and 0.34 for Salvini), probably because Movimento Cinque Stelle was openly critical of journalists, and Di Maio tends to overlap the interviewer in the discussion. The overall impression is that Di Maio has a good control over the conversation and does not let emotions interfere much with the flow of the debate. Also when he smiles or laughs, his body and eyes do not move much and are not used to emphasize a message.

This kind of control is even more evident in Matteo Salvini's interview. The only non-verbal devices he uses to convince the audience are smiles and hand movements, especially complex hand trajectories. The gaze, the eyes and the eyebrows do not move at all. As regards lexical and semi-lexical traits, he uses repetitions slightly more frequently than Renzi (0.22 per turn on average) and only few truncations (0.09 per turn). The overall impression he gives is that of a cold-blooded person who is in control of the situation, whose persuasion strategy relies on his seriousness, paired with the worried attitude for the future of the country that he expresses throughout his arguments.

For the records, Luigi di Maio and Matteo Salvini won the following elections and became the Minister of Economic Development and the Minister of the Interior respectively.

Chapter 4

Co-Gesture Analysis

A gesture is a visible action of any body part when used as an utterance or part of an utterance (Kendon, 2004). We can talk about co-speech gestures if such actions are produced while speaking. Their occurrence, simultaneous or concomitant to speech, has led to different views regarding their role in communication (Wagner, Malisz, and Kopp, 2014).

Some authors (McNeill, 2005; Kendon, 2004) have considered gestures an integrative, inseparable part of the language system since speaking itself is regarded as a variably multimodal phenomenon (Cienki and Müller, 2008). Indeed gestures may provide important information or significance to the accompanying speech and add clarity to discourse (Colletta et al., 2015); they can be employed to facilitate lexical retrieval and retain a turn in conversations (Stam and McCafferty, 2008) and assist in verbalizing semantic content (Hostetter, Alibali, and Kita, 2007). From this point of view, gestures facilitate speakers in coming up with the words they intend to say by sustaining the activation of a target word's semantic features long enough for the process of word production to occur (Morsella and Krauss, 2004).

Co-gesture speech can also refer to the spoken words or phrases that are co-produced with hand gestures in face-to-face spoken conversation (Lin, 2017). According to (Krauss, 1998) these co-occurring words or entire lexical phrases were identified to reflect the meaning of the co-occurring

gesture; they are also known as "lexical affiliates" of the gesture, especially if they play a particular role in the lexical retrieval. Indeed if gestures play a role in lexical retrieval, they must stand in a particular temporal relationship to the speech they are supposed to facilitate. For example, it would be difficult to argue that a gesture helped a speaker retrieve a word if the gesture was initiated after it had been uttered.

A 2015 study by Colletta et al. (Colletta et al., 2015) focused specifically on co-speech gesture production in children's narratives, language syntax influences gesture production. For example - as known - some languages require an explicit subject (i.e., English, French, etc.), whereas others (i.e., Italian, Spanish, etc.) are null-subject languages. This characteristic requires the distinct marking of referential continuity in the textual use of language, with less need to repeat anaphora in the latter case (Hickmann, 2002).

In addition to language differences, another critical factor influencing multimodal communication is culture as a set of values and norms that helps shape the social behavior of individuals who belong to a cultural group and the social interaction between them. Very well known is the study in (Kendon, 2004), showing that Italians use a significant number of gestures when communicating.

Whatever the case, co-speech gestures vary in different respects (Wagner, Malisz, and Kopp, 2014). Initially, (McNeill, 1992) differentiated them along *Kendon's continuum*. With a higher degree of conventionalization, the gesture becomes less dependent on the co-occurring speech, with sign language being completely independent. Emblematic gestures, e.g., the "thumbs up" gesture, are conventionalized and language-specific. By contrast, co-speech gestures are less standardized and work only in correlation to speech to accomplish communicative success. Later, (McNeill, 2005) refined the idea and argued for a complex of several continua, namely:

- Continuum 1: relationship to speech (from the obligatory presence of speech to obligatory absence of speech)
- Continuum 2: relationship to linguistic properties (from the absence of linguistic properties to the presence of linguistic properties)
- Continuum 3: relationship to conventions (from not conventionalized to fully conventionalized)
- Continuum 4: the character of semiosis (from global & synthetic to segmented & analytic)

Gesticulations are placed on the left ends of these continua (co-speech, no linguistic properties themselves, not conventionalized, global meaning).

As pointed out by (Lin, 2017) two gesture-speech characteristics can explain the link between speech and gesture: semantic coherence (combining gesture with meaningful and related speech) and temporal synchrony (producing gesture in synchrony with speech) (Butcher, 2000). The role of synchronization is particularly relevant for creating multimodal resources (Allwood, 2008). It allows researchers to overcome one of the historical limits of traditional one-modality corpora (written or spoken): presenting data in a single format offers limited opportunities for exploring non-verbal, gestural discourse features. At the same time, they are essential to understanding intercultural face-to-face interaction (Adolphs and Carter, 2013; Knight, 2011). Concerning the role of synchronization (Allwood, 2008) suggest adopting the general principle of spatiotemporal contiguity. It means a text occurs simultaneously with the event it describes or represents.

Therefore, when the two contiguous elements are transcribed speech and recorded speech (i.e., in audio-video format), we can more explicitly talk about of “synchronized alignment” of recording and transcription.

This alignment provides a comprehensive overview, allowing simultaneous visualization of the audio and video sources to which the annotation refers and vice versa.

4.1 Gesturing with hands

The gestural movements of the hands and arms, i.e., spontaneous communicative movements that accompany speech (McNeill, 2005), are probably the most studied co-speech gestures (Wagner, Malisz, and Kopp, 2014). According to the seminal works by (Kendon, 1972) about the relationship between body motion and speech and by (Kendon, 1980) about gesticulation and speech in the process of utterance, they are usually separated into several *gestural phases*: rest position, the preparation phase, gesture stroke, holds and retraction or recovery phase (Bressemer and Ladewig, 2011). Additionally, the maximal gestural excursion point is often regarded as a gestural *apex*.

More generally, gestures can be described in terms of their form, semantic and pragmatic functions, temporal relation with other modalities, and relationship to discourse and dialogue context. One of the most well-known classifications is certainly that of (McNeill, 1992), which attributes five semiotic functions to hand movements:

- *emblematic gestures* bear a conventionalized meaning (“thumbs up”);
- *iconic gestures* resemble a specific physical aspect of the conveyed information, i.e. they may convey the shape of a described object or the direction of a movement;
- *metaphoric gestures* are iconic gestures that resemble abstract content rather than concrete entities (McNeill, 1992; Cienki and Müller, 2008);

- *beat gestures* are simple and fast movements of the hands (also called batons (Ekman and Friesen, 1972)).

This classification should not be understood as defining distinct categories. (McNeill, 2005) argued that a simple, functional classification of gestures is usually misleading. By contrast, (Wagner, Malisz, and Kopp, 2014) proposes a more nuanced classification that considers the multifaceted nature of most gestures. Gestures are classified using dimensional criteria, such as iconicity, metaphoricity, deixis, temporal highlighting (beats), and social interactivity. These dimensions can characterize the majority of gestures, i.e., when a pointing gesture also depicts the direction of a movement or when a beat is superimposed onto the stroke onset of an emblematic gesture (Tuite, 1993).

A further classification is that proposed by (Lin, 2017) adapting (Colletta et al., 2015; Kendon, 2004), according to which movements (of the hands in particular) assume five possible functions:

- *reinforcing*: the information the gesture brings equals the linguistic information it is related to. For example, the gesture can resemble the physical properties and movement of objects or actions described in speech. For example, in the corpus analyzed by us, one of the interviewees emphasizes the sacrifices to which Italians have been subjected in the last fifteen years, including "il 3% del rapporto deficit/PIL (*en.* the 3% deficit/PIL ratio)". In saying this, he makes the sign of the number three with the fingers of his right hand.
- *integrating*: the information provided by the gesture does not add supplementary information to the verbal message but makes the abstract concepts more precise. A frequent example in our dataset is respondents who, in order to contrast two items, such as left and right parties, point their hands one toward the right and the other toward the left.

- *supplementary*: the information brought by the gesture adds new information not coded in the linguistic content. In one of the interviews we analyzed, the interviewee regarding the number of politicians that a rival would have brought to Parliament says, "...non so quanti parlamentari porterà in Parlamento" (*en.* ...I don't know how many parliamentarians he will bring to Parliament" and in the meantime, he opens his arms as if to imply a large number.
- *complementary*: the information provided by the gesture brings a necessary complement to the incomplete linguistic information provided by the verbal message. The gesture usually disambiguates the message, for example, in our dataset it is common to find cases where deictic adverbs such as *qui* (*en.* here) are accompanied by the corresponding pointing gesture.
- *contradictory*: the information provided by the gesture contradicts the linguistic information provided by the verbal message. This type of gesture has never been found in our dataset. However, an example is provided by the dataset of (Lin, 2017) where a Taiwanese speaker was saying comes to Taiwan, a gesture with two opening hands moving toward the hearer was produced.
- *other*: within this category that we could define - as in the social sciences - residual, fall all those gestures that the annotator was not able to classify with the above-mentioned semantic labels.

Since the above classification can effectively capture the semantic contribution of gestures w.r.t. the (written or oral) utterances, we adopt it in our study. We include such classes in our classification scheme.

In particular, our annotation follows the selection criterion highlighted by (Allwood et al., 2007). Annotators will keep track only of gestures they believe have a communicative function concerning the speech produced (Allwood, 2001) or judged to have a noticeable effect on the recipient.

However, as (Yoshioka, 2008) points out gestures can be functionally ambiguous and thus have multiple semantic functions simultaneously. According to (Tsui, 1994), the source of this multiple functions often lies in the sequential environment of the conversation in which the utterance occurs. In this case study, we try to attribute a single semantic function to the gestures under investigation with respect to the function we considered primary to the context of use.

Starting from research questions introduced in the section 1.1, this work aims to prove how the use of multi-level annotated multimodal corpora can improve the quantitative and qualitative study of the co-occurrences between gesture and speech, focusing on the Italian language, hitherto little considered in the literature for this kind of studies.

4.2 Coding co-speech gesture

The video annotation has been carried out using the ANVIL tool (Kipp, 2001). Levels and labels used in the annotation scheme (which includes hand movements) combine the MUMIN coding scheme notation (Allwood et al., 2007) for the typology and physiology of the gestures to be annotated and - as specified above - the classification proposed by (Lin, 2017) adapting (Colletta et al., 2015; Kendon, 2004) for the semantic function of the movement.

The written-spoken alignment facilitated annotating hand movements produced at particular utterances. As specified earlier, this study focuses mainly on co-speech hand gestures, which - according to (McNeill, 2005) - can be defined as spontaneous communicative hand movements that accompany speech.

So at the end of the annotation process, the xml markup is as follows:

The annotation task addressed so far falls – from a qualitative point of view – in the first general types identified by (Mathet, Widlöcher, and

```

<u gender="m" length="928" role="Minister of the foreign
business and of the international cooperation" time=
"452.28" who="Angelino Alfano">C'è qualcosa di più grave
e di più profondo di cui mi sono occupato da Ministro
dell'Interno. Perché io ho gestito l'immigrazione
<movement start="470.2" end="471.2" attribute="Hand
movement trajectory" attribute_text="sideways" function=
"integrating">e ho gestito l'ordine pubblico.
</movement></u>

```

FIGURE 4.1: Annotation extract in xml

Métivier, 2015), in which the subjective interpretation is limited. Indeed, it deals with the "identification of units" (Krippendorff, 2018), in which the annotator, given a written or spoken text, must identify the position and boundary of linguistic elements (i.e., identification of prosodic or gestural units, topic segmentation).

The annotation – made at the beginning of the process by a single expert annotator – follows the criterion highlighted on the one hand by (Allwood et al., 2007) and on the other hand by (Kendon, 2004).

According to Allwood, annotators only consider gestures with a communicative function. As regards the annotation guidelines - as specified in (Allwood et al., 2007) - the attributes concerning the shape or dynamics of the observed phenomena are coarse-grained because they only seek to capture features that are significant for the functional level of the annotation. Once an annotator selects a gesture, he performs the modality-specific annotation level, labeling the gesture with functional values and features that describe its behavioral shape and dynamics.

While during the process of coding co-speech gestures, according to Kendon, the coders must consider three criteria in particular: if the movement is easy to perceive, of good amplitude, or marked well by its speed; if the location of the movement is in frontal space of the speaker; if there is a precise hand shape or a well-marked trajectory.

Subsequently, to evaluate the reliability of our annotation scheme, we

compute inter-annotator agreement double annotating verbal and non-verbal traits of the first ten minutes of Renzi's, Di Maio's, and Salvini's interview. Both annotators were expert linguists. Macro-averaged F1 computed on exact matches amounts to 0.82, which corresponds to a good agreement, given that by exact match, we consider the correct choice of the trait, the position of the tag, and the exact extension of the marked string, if any. This result confirms the reliability of the task and the corresponding annotation guidelines.

The decision to focus on a small sample of analysis to study the correlation between particular semantic types of gestures and speech is motivated by the high cost in terms of time and computation involved in the process of annotation and revision.

4.3 Which type of verb do hand movements accompany most frequently?

The study presented in (Vignozzi, 2019) aimed at analyzing the representation of some peculiar indicators of spokenness (i.e., idiomatic expressions, and phrasal verbs) across TV interviews featuring different interviewees (politicians, business people, and personalities from showbiz). The analysis pointed out that phrasal verbs are more recurrent in political interviews than business and economic discussions and that the specialized context with which hand or arm movements are more often associated is the business and economics domain (60.86%). In political interviews, instead, gestures appear in 58.02% of cases, while in television interviews, the lowest frequency is observed since gestures occur only in 40.42% of the cases. Besides, the study shows that beats gestures are the most frequent ones co-occurring with phrasal verbs, especially in political interviews, which account for more than half of the total gestures. The study was conducted on "The ESP Video Clip Corpus" in English.

Recent research on multimodal corpora in Italian is limited to the experience of the IMAGACT project (Moneglia et al., 2014). IMAGACT aims to set up a cross-linguistic Ontology of Action for grounding disambiguation tasks and uses the universal language of images to identify action types, avoiding the under-determinacy of semantic definitions. As far as it is known, no other studies have been identified focusing on the co-presence of gestures and particular types of verbs in Italian, so this experimentation represents a first step in this direction.

In order to understand if the tag *hand movement trajectory* occurs systematically with particular verbal modes and tenses and/or defined verbal types (i.e., movement verbs, phrasal verbs etc.), only sentences co-occurring with the tag under investigation were extracted (for a total of 495 tags). These were post-tagged in such a way as to be able to identify verbal tenses and modes, using TINT (Palmero Aprosio and Moretti, 2018). Finally, they were subject to a qualitative analysis to identify particular verb types.

The qualitative approach has been preferred in this phase for two main reasons: first of all, because the amount of data to be analyzed is controllable; moreover, because existing resources for Italian such as LexIt (Lenci, Lapesa, and Bonansinga, 2012), MultiWordNet (Pianta, Bentivogli, and Girardi, 2002) and T-PAS (Jezek et al., 2014) do not make explicit the function that the verbs assume in the context (i.e., no tool will tell us if the verb is servile, appellative, estimative, elective, etc.). This is what has been done by the Lexicon-Grammar (LG) (Gross, 1994) that can group verbs into classes according to their behavior, independently of the language used. LG tables for Italian include about 7000 entries (D'Agostino, Elia, and Vietri, 2004) annotated in detail by expert linguists. Despite this large amount of available information, LG tables have some limitations

that have made them rarely usable in the recent literature. In particular, they are verbose, and the structure of the properties is neither uniform nor standardized. Furthermore, the distributional properties are not directly usable in a computational approach since they include semantic constraints (Guarasci et al., 2020).

An additional factor of difficulty in classifying Italian verbs can probably be attributed to the complexity of the verb inflection (as it happens in other Romance languages such as French). However, some information about verbs can be extracted using automatic text analysis techniques. For instance, using morphological analysis, we can quickly identify the mode (finite or indefinite), the tense (present, past etc.), and the person (first, second, and third, singular or plural) of the verb. However, these analyses are not helpful in understanding function or role of verbs in the context, so an additional level of qualitative analysis is needed.

In this first exploratory phase, no distinction was made between the single interviews, but instead, a sub-corpus was created containing only the sentences co-occurring with the tag of interest (hand movements) for a total of 495 sentences. The results obtained from the automatic analysis show the presence of 4,209 verbs, of which 823 auxiliary verbs. An overview of the frequencies of each verb mode in the sub-corpus is reported in Table 4.1.

Among the verbal modes, there is an evident prevalence of the indicative (1,989), used to talk about what is or what we consider trustworthy and safe. In the interviews under examination, it is mainly used to underline the interviewees' awareness. In some cases, the sentences are accompanied by gestures with an integration function in order to make abstract concepts more precise. See the examples below:

- (1) *Veltroni: Ha visto la Repubblica Ceca, **sappiamo** cosa succede in Ungheria. In Francia la Le Pen ha preso il 30% in Germania l'FD è il terzo partito, in Austria il secondo partito.* (En. He saw the Czech Republic, what is

Verb Mode	Frequency
Ind	1989
Sub	98
Cnd	18
Imp	18
Ger	97
Inf	607
Part	647

TABLE 4.1: Absolute frequencies of verbal modes

happening in Hungary. In France Le Pen took 30% in Germany the FD is the third party, in Austria the second party)

- (2) *Di Stefano: Ovviamente **chi utilizza** la foto di Anna Frank pensando di offendere qualcuno è fuori, fuori di testa ovviamente ma io da romanista non mi sento offeso nel vedere...* (En. Obviously those who use the picture of Anne Frank thinking that they are offending someone are out, out of their minds obviously but I as a Romanist am not offended in seeing...)

This behavior is not surprising since in television interviews, the main objective is the ‘overhearing audience’ (Heritage, 1985), i.e. a person who, while not actively participating in the conversation, assumes the role of representative of the entire audience to whom the interview is addressed. Similarly, the interview is conceived as a triangular relationship between the interviewer, interviewee, and audience, shaping how knowledge is constructed during the interview (Furkó and Abuczki, 2014).

Concerning the linguistic register used during political interviews, past studies (Fairclough, 1998; Bruti, 2016) have defined these interactions as

stylistically hybrid, mixing elements typical of different registers (e.g., formal, institutional, informal, colloquial) and discourses (e.g., planned, unplanned). More generally, the process by which language undergoes adaptation for the audience's benefit inclines toward what is known as 'conversationalization of discourse' (Fairclough, 1998; Fairclough, 2000). This implies that the style becomes increasingly colloquial, involving emotive and more subjective linguistic strategies that help build rapport between interlocutors (direct or indirect) (Fairclough, 1998; Fairclough, 2000).

Among the finite modes, we observe a prevalence of the participle (647). In the sub-corpus considered, the participle acquires the function of an adjective in most cases. See for example the following excerpt:

- (3) *Alfano: E poi cosa si è verificato? Che nell'Aprile del 2016, quindi 8 mesi prima che io andassi via, è arrivato a Tripoli non con la banda musicale, con la fanfara ma con un barcone **proveniente** dalla Tunisia, il primo Ministro al-Sarraj, quello dell'Onu.* (En. And then what occurred? That in April 2016, so eight months before I left, he arrived in Tripoli not with the marching band, with the fanfare but with a barge from Tunisia, Prime Minister al-Sarraj, the one from the OUN.)

Sentences that co-occur with the tag under analysis present 331 first-person singular verbs; 1,134 verbs in the third-person singular; 160 first-person plural verbs; 326 third person plural verbs. The third person singular is predominant because it is used to address the interviewer-journalist with whom a well-known formal person is clearly used.

If the situation had been informal, probably the sender would refer to himself with *Io* (I) (first person) and to the addressee with *Tu* (You). In that case, the first and second person would therefore designate real persons, while the third person would be exclusively used to refer to "[...] an 'entity' that is not necessarily present and does not even need to be a 'person'" (Simone, 1990).

Through a qualitative analysis, we then manually classified verbs according to their function in the text. The verbal classes identified are as follows (with the total number of occurrences in parenthesis):

- Predicative verbs: they have full lexical meaning and can independently give rise to a verbal predicate of full meaning. The class of predicative verbs encompasses the vast majority of verbs in a language and is descriptively opposed to the class of copulative verbs that need to rely on a predicative complement to fulfill the predicate function: *sembrare* (13) (*to seem*), *parere* (*to look like*) (5), *risultare* (*to prove*) (4), ***stare (to stay)*** (131), *restare* (*to stay*) (7), *rimanere* (*to stay*) (2), *diventare* (*to become*) (5), *divenire* (*to become*) (0)
- Predicative verbs which can carry a predicative complement of the subject, but only if conjugated in the passive form distinguished in:
 - appellativi: *chiamare* (2), *definire* (0), *denominare* (0), *appellare* (0)
 - elettivi: *eleggere* (2), *nominare* (0), *proclamare* (0)
 - estimativi: *stimare* (0), *giudicare* (1), *ritenere* (0), *considerare* (0), *reputare* (0)
 - effettivi: ***fare*** (12) and *rendere* (0)
- Phrasal verbs are verbs that, when combined with another non-finite mode verb with the interposition of a preposition (*to*, *of*, *for*, *from*), specify a particular time-expectant mode. They are divided into 5 groups:
 - the imminence of an action: *stare per* (3), *accingersi a* (0), *essere sul punto di* (0), *stare lì lì* (0) + infinitive
 - the beginning of an action: *cominciare a* (7), *mettersi a* (0) and *prendere a* (0) + infinitive

- the development of an action: *stare* (38), *andare* (0) and *venire* (15) + *gerund*
- the duration and continuity of an action: *continuare a* (6), *insistere a* (0) and *ostinarsi a* (0) + infinitive
- the conclusion of an action: *finire di* (1), *cessare di* (0) and *smettere di* (1) + infinitive
- Causative verbs: indicate that the subject causes the action but that he does not perform it directly. The two causatives of the Italian language are *fare* (20) and *lasciare* (0) + infinitive
- Performative verbs exist only in the first person singular of the present indicative. They are so defined because pronouncing them is equivalent to performing the action they describe. The verbs most representative of this category are *giurare* (0), *promettere* (0) and *negare* (1).¹ Other verbs that in the first person of the present indicative take on a performative function are, for example: *dire* (26), *ammettere* (0), *affermare* (0), etc.

Most function verbs are predicative: they have an independent meaning, forming what in syntax is called a verbal predicate. Among them, we notice the more frequent use of the verb **stare** (to stay) with 131 occurrences.

- (4) *Salvini: Ci possono essere altre sfumature, a qualcuno **sta** simpatico Macron, a qualcuno **sta** simpatica la Le Pen, è il rapporto con l'Europa che per me è determinante al di là delle simpatie.* (En. There may be other nuances, someone likes Macron, someone likes Le Pen, it is the relationship with Europe that for me is decisive beyond sympathies.)

¹See for example the utterance by Di Battista: *Io ho avuto credo 84 giorni di espulsione dalla Camera dei Deputati e non ho mai picchiato nessuno, mai. Anche se non le nego...;* En: (I think I had 84 days of expulsion from the Camera dei Deputati, and I never beat anybody, ever. Even though I don't deny...)

Among verbs with a predicative function of the subject (only when used in the passive form), the most commonly used are effective verbs, i.e. copulative verbs indicating a state, semblance, or transformation. In this case the most frequent is **fare** (to do) with 12 occurrences.

- (5) *Padoan: Secondo te questa campagna elettorale sta dividendo il paese in due. Tra chi vuole continuare e rafforzare quello che è stato fatto e ha portato i risultati che lei ricordava, piuttosto che chi vuole eliminare.* (En. According to you, this election campaign is dividing the country in two. Between those who want to continue and strengthen what has been done has brought the results you mentioned, rather than those who want to eliminate.)

On the other hand, concerning phrasal verbs, the results obtained do not confirm what emerged in (Vignozzi, 2019), in which a predominance of servile verbs was noted in political domain interviews. In our case, there is a slight but not apparent prevalence of verbs that indicate the performance of an action, in particular of the verb **stare** (to stay) + gerund with 38 occurrences.

- (6) *Veltroni: E quello che sta succedendo in Italia, l'affermazione non delle forze tradizionali...* (En. And what is happening in Italy, the assertion not of traditional forces...)

Among causative verbs, the most present is the verb **fare** (to do) (20 occurrences), while among performative ones, it is **dire** (to say) (26).

- (7) *Tremonti: E quando comincio a vedere che perfino Prodi parla di un colpo di quel tipo, avremmo dovuto andare a votare e non ci hanno fatto andare a votare. Perché dovevano mandarci il Governo tecnico che tecnicamente ci ha buttato giù.* (En. And when I start to see that even Prodi is talking about that kind of blow, we should have gone to vote and they didn't let us go to vote. Because they had to send us the technical government that technically brought us down.)

- (8) *Di Maio: Guardi io le dico noi parleremo con tutti coloro che aderiranno però...* (En. Look I tell you we will talk to everyone who will adhere but...)

Causative verbs (also called factitive verbs) express an action not performed by the subject but made to be performed by others. In this case, we notice a prevalence of the verb *fare*, mainly used with negative valence and referred to the political opposition; in fact, this verb mainly describes actions the subjects were forced to carry out because of the determined political circumstance of the moment.

The theory of linguistic acts elaborated in (Austin, 1975) introduced the concept of performative act concept. Verbs that take on this function are so defined because pronouncing them is equivalent to performing the action they describe. In other words, to perform the action they describe, one must pronounce them. Probably the performative verb *dire* is more present in these interviews because politicians in the middle of an electoral campaign want to give an impression of being concrete and aim at emphasizing their statements.

4.4 Is the Lexical Retrieval hypothesis confirmed?

Many studies have suggested that gestures, especially representational gestures (Krauss and Hadar, 1999), play a direct role in speech production through priming the lexical retrieval of words. So as mentioned in the previous paragraphs, this view has been termed the *Lexical Retrieval hypothesis*. The hypothesis was based on research arguing that (1) gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978), and (2) that the inability to gesture can cause verbal disfluencies (Dobrogaev, 1929). In addition - as (Krauss, 1998) pointed out

- speakers were more dysfluent overall in the obscure and constrained-speech conditions than in the natural condition.

Since the corpus used as the object of study presents a level of annotation that takes into account some hesitation pauses and verbal disfluencies, besides a level of proxemic annotation (which therefore takes into account the more strictly gestural part) we decided to verify the *Lexical Retrieval hypothesis* (LR, henceforth) specifically in the political domain.

Before describing the analyses conducted and the methodology used, we introduce some theoretical clarifications on the elements the LR considers.

4.4.1 Background to Lexical Retrieval Hypothesis

Speakers differ very much in their rhythm, particularly in the time left between words or utterances. Regarding *hesitation pause*, several studies have concluded that we alternate between planning speech and implementing our plans. Indeed, as shown in (Henderson, Goldman-Eisler, and Skarbek, 1966), interview participants typically show a cycle of hesitation and fluency, although the ratio of speech to silence varies among speakers. The pause tag is therefore used to mark when the speech has been paused, irrespective of the actual amount of silence. A pause marked within an utterance applies to the speaker of that utterance.

A typical aspect of spoken language is the use of *disfluencies* (Voghera, 2001). Dysfluencies have been classified in several ways in the literature, including silent and filled pauses, truncated and repeated words (Krauss, 1998), interjections, and false starts.

Interjections (e.g. ah, ehm) have been associated with the fact that linguistic planning is very cognitively demanding, and it is difficult to plan an entire utterance at once (Lindsley, 1975). Therefore, hesitation pauses

and similar vocal phenomena may be helpful to to perform a careful lexical retrieval since past studies (Levelt, 1983) found that pauses occurred more often before low-frequency words than before high-frequency ones.

Although spoken texts are the product of a physically continuous process, their structure shows a strong discontinuity: *false starts*, interruptions, and project changes are common to all spontaneous speech texts. In particular, by false start we mean the abandonment by the speaker of a word or a sequence of words already produced, with or without repetition of the previously used linguistic material (Cresti and Panunzi, 2013).

The spontaneous speech presents a frequent and somewhat ‘compulsory’ use of deictic elements (Givón, 1995). Some forms of ellipses can also be traced back to deictic phenomena (Berretta, 1994). The same need for indexicality can also be traced back to cases of reduction and *truncation*.

4.4.2 Verification of the Lexical Retrieval Hypothesis

In order to verify the LR, we computed the association between hand movements produced by each interviewee and speech disfluencies by calculating the *weighted mutual information* (WMI, a form of weighted KL-Divergence, which is known to take negative values for some inputs (Kvålseth, 1991). There are examples where the weighted mutual information also takes negative values (Pocock, 2012).

In the traditional formulation of mutual information:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x) p(y)}$$

FIGURE 4.2: Traditional formula of Mutual Information

each event or object specified by (x, y) is weighted by the corresponding probability $p(x, y)$. This assumes that all objects or events are equivalent apart from their probability of occurrence. However, in some applications,

certain objects or events may be more significant than others, or certain patterns of association are more semantically important than others.

For example, the deterministic mapping $[(1,1),(2,2),(3,3)]$ may be viewed as stronger than the deterministic mapping $[(1,3),(2,1),(3,2)]$ although these relationships would yield the same mutual information. It is because the mutual information is not sensitive to any inherent ordering in the variable values (Cronbach, 1955; Coombs, Dawes, and Tversky, 1970; Lockhead, 1970) and is therefore not sensitive at all to the form of the relational mapping between the associated variables. If it is desired that the former relation—showing agreement on all variable values—be judged stronger than the later relation, then it is possible to use the following weighted mutual information (Guaşu, 1977).

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} w(x, y) p(x, y) \log \frac{p(x, y)}{p(x) p(y)}$$

FIGURE 4.3: Weighted Mutual Information formula

which places a weight $w(x,y)$ on the probability of each variable value co-occurrence, $p(x,y)$. It allows certain probabilities to carry more or less significance than others, thereby allowing the quantification of relevant holistic or pragmatic factors. In the above example, using larger relative weights for $w(1,1)$, $w(2,2)$ and $w(3,3)$ would have the effect of assessing greater informativeness for the relation $[(1,1),(2,2),(3,3)]$ than for the relation $[(1,3),(2,1),(3,2)]$, which may be desirable in some cases of pattern recognition, and the like.

Chapter 5

Results and Discussion

Given the purpose of this analysis, we consider the interviews in our corpus that have a minimal length of 50 turns.

In the following table we present the normalized frequencies of different phenomena annotated in our corpus for each politician under analysis. For each phenomenon we compute the incidence per 100 turns.

Interviewee	Hand mov.	Pause	Semi-Lexical	False-start	Repetition	Truncation
Matteo Renzi	35.82	0	8.5	10.16	22.45	36.89
Luigi Di Maio	22.97	0	14.86	0	18.91	18.91
Matteo Salvini ₁	54.38	5.2	24.56	0	24.56	19.29
Matteo Salvini ₂	52.87	14.94	21.83	3.44	21.83	3.44
Walter Veltroni	41.81	0	14.54	21.81	29.09	18.18
Simone Di Stefano	10.98	0	4.39	5.49	21.97	16.48
Pierluigi Bersani	32.29	1.04	26.04	0	31.25	20.83
Angelino Alfano	57	9	33	3	17	3
Giulio Tremonti	10.71	16.07	10.71	0	14.28	0
Matteo Orfini	29.85	1.49	11.94	0	14.92	0
Pier Carlo Padoan	49.27	11.94	30.43	1.44	7.24	13.5
Carlo Calenda	74.63	32.60	24.63	9.42	7.24	0.72
Alessandro Di Battista	39.02	9.26	32.19	6.82	11.70	10.58

TABLE 5.1: Normalized frequencies of the tags for each interview

Among the politicians considered in this dataset, the one that most accompanies his speech with the movements of the hands is Matteo Salvini

(Lega) [107.25, considering both interviews] followed by Carlo Calenda (PD) [74.63] and Angelino Alfano (Il Popolo della Libertà) [57].

If we compare the normalized frequencies of hand movements with respect to the political party to which they belong, we can easily note that - with the exception of the Popolo delle Libertà - represented by only one interviewee (Angelino Alfano), the representatives of the PD (Matteo Renzi, Carlo Calenda, Pierluigi Bersani, Walter Veltroni, Pier Carlo Padoan, Matteo Orfini) are those who make greater use of their hands to accompany speech, especially to strengthen what they say. They are followed by Lega (Matteo Salvini, Giulio Tremonti), M5S (Luigi Di Maio, Alessandro Di Battista) and Casa Pound (Simone Di Stefano). The outcome of this party-based comparison is reported in Figure 5.1.

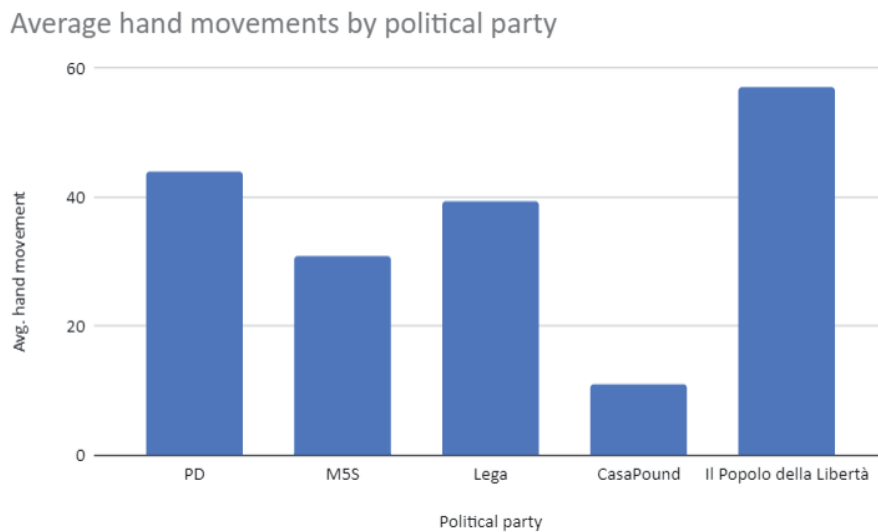


FIGURE 5.1: Avg. hand movements per political party

In order to calculate WMI, we first count how many times hand movements occur in correspondence with the disfluences under examination. Then, by applying the above formula, we calculate WMI, with the goal to

capture whether co-occurrence happens by chance or is significantly more frequent with some traits.

Weighted mutual information with Hand Movement

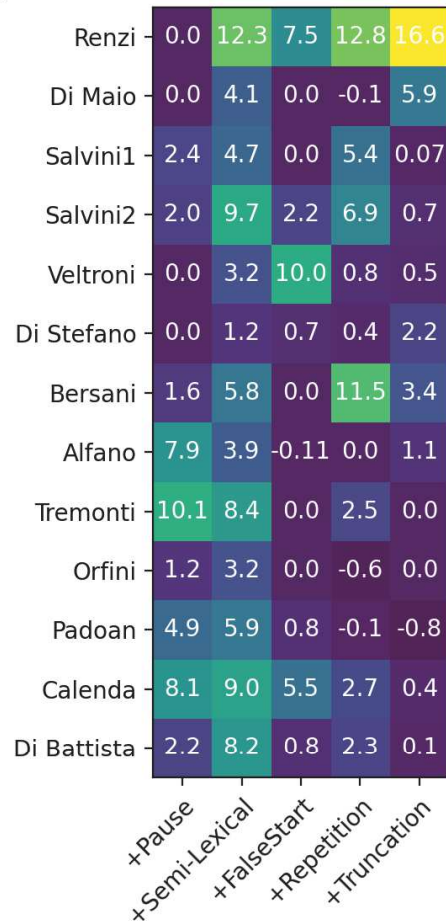


FIGURE 5.2: WMI values for each tag divided by interviewee

In order to understand the correlation of hand movements with all the traits considered, the average WMI was calculated for each traits which are respectively:

- pause: 3,10

- semi-lexical: 6,12
- false start: 2,10
- repetition: 3,42
- truncation: 2,30

The average values obtained suggest that – in this specific case study – hand movements tend to co-occur with full pauses (i.e. repetition) and empty pauses (i.e. pause) and more frequently with interjections (i.e. semi-lexical), suggesting that gesticulating may represent an attempt at lexical retrieval. As suggested by literature (Lindsley, 1975) these traits have been associated with the fact that linguistic planning is very cognitively demanding, and it is difficult to plan an entire utterance at once. Therefore hesitation pauses and similar vocal phenomena may be useful to perform a careful lexical retrieval, since pauses occurred more often before low-frequency words than before high frequency ones (Levelt, 1983). This would confirm the assumptions of *Lexical Retrieval hypothesis* according to which gesturing occurs during hesitation pauses or in pauses before words indicating problems with lexical retrieval (Dittmann and Llewellyn, 1969; Butterworth and Beattie, 1978). This effect is however not present for some politicians, such as Di Battista and Alfano, while it is evident for some others such as Bersani and Salvini. Therefore, our findings are not generally applicable to all interviewees in our corpus.

In the individual interviews, the negative values obtained in relation to false-starts (-0,11), repetitions (-0,1 and -0,6) and truncations (-0,8) suggest that hand movements are less likely to be accompanied by such linguistic phenomena, which in any case are not properly associated with lexical recovery problems. Indeed, several past studies (Simone, 1990; Tanen, 1989; Bazzanella, 1992) highlighted their importance in spontaneous speech, but mentioned in particular the role of repetitions in controlling

5.1. Is the gesture-speech relationship influenced by linguistic variables⁶⁷

the in-progress textual design of speech (Voghera, 2001). False start and truncations tend to appear *less frequently* in association with hand movements.

5.1 Is the gesture-speech relationship influenced by linguistic variables?

Finally an analysis was carried out in order to understand if the hand movements produced by the interviewees have significant correlations with language complexity. As in the previous analysis, the only focus was on the interviews that have a minimal length of 50 turns.

For complexity we consider type-token ratio and conceptual density. We perform an analysis of the correlation between language complexity and hand movements automatically annotated with the ANVIL software, normalised by the number of tokens uttered by each politician multiplied by one thousand.

Since the variables under examination are both cardinal or quantitative, the Person's correlation index had been used for each interviewee and for each political party they belong to.

Individual interviewee computations reveal that both the TTR and the conceptual density show a moderate negative correlation with hand movements, respectively $r = -0,3$ and $r = -0,12$.

Since in all cases considered the correlation is negative it could deduce that the Information Retrieval hypothesis is confirmed.

The value of the TTR could mean that the more you gesticulate the more the lexical richness decreases and therefore there are more hesitations. Instead in the case of conceptual density, the negative value $r = -0,12$ could mean that the more you gesticulate the more the speech tends to be simple and understandable (this could find even more justification in the format of the interview that being televised and being broadcast at

Interviewee	Hand movement	TTR	Lexical Density
Matteo Renzi	35.82	0,71	0.563
Luigi Di Maio	22.97	0,8	0.562
Matteo Salvini ₁	54.38	0,73	0.567
Matteo Salvini ₂	52.87	0,82	0.569
Walter Veltroni	41.81	0,7	0.569
Simone Di Stefano	10.98	0,75	0.583
Pierluigi Bersani	32.29	0,73	0.547
Angelino Alfano	57	0,61	0.564
Giulio Tremonti	10.71	0.75	0.585
Matteo Orfini	29.85	0.72	0.566
Pier Carlo Padoan	49.27	0.75	0.570
Carlo Calenda	74.63	0.73	0.580
Alessandro Di Battista	39.02	0.8	0.568

TABLE 5.2: Normalized values of hand movements, TTR, and lexical density for each interviewee

a time when the audience is quite varied, it could tend to be easier to be understood by all).

Political Party	Avg. Hand movement	Avg. TTR	Avg. Lexical Density
PD	43.94	0.73	0.566
M5S	30.99	0.80	0.565
Lega	39.32	0.76	0.574
CasaPound	10.98	0.75	0.583
Il Popolo della Libertà	57	0.61	0.564

TABLE 5.3: Values of hand movements, TTR, and lexical density for each political party

Also political parties computations reveal that both the TTR and the conceptual density show a moderate negative correlation with hand movements, respectively $r = -0,7$ and $r = -0,71$ even if slightly higher than the correlation per single respondent with a deviation of 0,5 for TTR and 0,6 for conceptual density.

The correlation values obtained by political party of belonging show a slight negative correlation, which could mean that the party of belonging does not significantly influence the use of the proxemic communication plan and consequently the use of language.

5.2 Semantic gesture-speech relationship

The last analysis performed was aimed to understand the nature of the relationship between gestures and speeches, to do this - after the identification of hand movements performed in correspondence with certain utterances - the gestures of interest were given a label indicating their function with respect to the reference sentences (*reinforcing, integrating, supplementary, complementary, contradictory, other*).

In the following table, we report the frequency of the type of gestures produced by each respondent interviewee and immediately following the results aggregated by political party.

From an initial observation of the results obtained from each individual interviewee, it appears clear that of 12 politicians, 10 (with the exception of Di Maio and Orfini) use hand movements with an integrative function (*integrating*). The information provided by such gestures adds precision to the abstract concepts of the linguistic information.

Below are some examples that exemplify how such gestures are used by politicians belonging to three different parties:

Matteo Renzi: Però, giusto per non perdere l'abitudine, non è che sia d'accordissimo sulla lettura che lei dà, nel senso che il referendum l'ho perso io. Non è che c'era il sistema contro. Io quei politici che danno la responsabilità, la colpa, si nascondono dietro gli alibi personalmente non li sopporto. Quindi non do responsabilità ad altri, quel referendum non è stato colpa del sistema, punto. Quello che sta accadendo invece in queste settimane, in questi mesi, conferma che c'è una grande distanza tra la politica dei palazzi e la politica della

Interviewee	int.	reinf.	suppl.	compl.	contr.	other
Matteo Renzi	32	9	2	23	0	1
Luigi Di Maio	6	0	1	9	0	1
Matteo Salvini ₁	16	6	3	5	0	1
Matteo Salvini ₂	17	10	0	14	0	5
Walter Veltroni	8	3	0	8	0	4
Simone Di Stefano	5	0	2	3	0	0
Pierluigi Bersani	13	4	0	12	0	2
Angelino Alfano	21	11	0	16	0	8
Giulio Tremonti	3	1	1	1	0	0
Matteo Orfini	7	0	0	10	0	3
Pier Carlo Padoan	16	0	0	3	0	15
Carlo Calenda	41	1	0	35	0	26
Alessandro Di Battista	29	1	0	20	0	0

TABLE 5.4: Frequency of the type of gestures produced by each interviewee

quotidianità [integrating]. Lo dico senza il tono populista o, se vuole anche un po' superficiale. Cioè io ho fatto il Presidente del Consiglio, sono il segretario del primo partito, non è che posso...

In this turn the interviewee responds to a reading made by the interviewer (Lucia Annunziata) regarding the motivations that would have led to the defeat of the constitutional referendum of December 4, 2016. The reform on the ballot - proposed by the then President of the Council of Ministers Matteo Renzi and the then Minister for Constitutional Reforms and Relations with Parliament Maria Elena Boschi - contained provisions for the overcoming of equal bicameralism, the reduction of the number of parliamentarians, the containment of operating costs of institutions, the abolition of the CNEL and the revision of Title V of Part II of the Constitution. In fact, the interviewee reads the defeat of the referendum and the political climate of that period as proof of the existence of a distance between the politics of the palaces and that of everyday life. This expression probably paraphrases the "metaphor of the Palace" coined in "Lettere

Political party	int.	reinf.	suppl.	compl.	contr.	other
PD	117	17	2	91	0	51
M5S	35	1	1	29	0	1
Lega	36	17	4	20	0	6
Casa Pound	5	0	2	3	0	0
Il Popolo delle Libertà	21	11	1	16	0	8

TABLE 5.5: Frequency of the type of gestures for each political party

Luterane" by the Italian poet, writer and director Pier Paolo Pasolini, who had addressed the theme of the split between politics and life by alluding to the physical space in which the former is exercised in Italy. In this case, Renzi underlines that the distance between an autonomous politics, separate and not very attentive to the real problems of the country (politics of the Palaces) and a politics of everyday life, that is, attentive to reality and to the citizens for whom it is exercised and has the right to exist, is increasingly evident.

This metaphorical expression is made more concrete by Renzi's hand movements. His open right hand points away from his torso in correspondence with the metaphorical expression "politica dei Palazzi", almost as if to indicate that it is something distant yet present but in which he does not recognize himself; his right hand then immediately rejoins his left hand and points downwards at the moment in which the expression "politica della quotidianità" is pronounced, as if to indicate a politics that is instead attentive to real, concrete and present things. Therefore metaphorical and apparently abstract concepts are concretized through the use of this gesture that therefore carries out a function of integration.

The second example we report has a simpler and more immediate reading. The interviewee speaks about the alliances necessary to bring the party to which he belongs to victory in the upcoming local elections of

2017. Referring to the percentages achievable by the center-right, the interviewee explains that his only request is the presentation of a common program. In pronouncing this expression, the fingers of the hands intertwine in order to make this concept concrete and readable. This gesture is easy to read because it is commonly used by Italian speakers to make this concept.

*Matteo Salvini: Il centro-destra può arrivarci al quaranta per cento e io chiedo al centro-destra che ci sia **un programma comune** [integrating] che prevede che a Bruxelles prevalga l'interesse nazionale italiano.*

In the third example, the respondent within the same dialogic turn produces three gestures with integrative function. In the sentences we are going to use, the interviewee explains the reason why in his opinion it is not true that a form of fascism is reborn in the center-right, and he does so by explaining his experience as Minister of the Interior. The first integrative gesture is produced in correspondence of the expressions "immigration and public order", two topics that the Minister says he has dealt with. These issues, belonging to two different fronts, are made explicit with both hands pointing first to the left and then to the right, as if to indicate that they are two different issues, belonging to two different levels but nevertheless addressed. The second gesture of the same nature is produced in conjunction with the discussion on the management of the immigration issue that the Minister says: "Which not even the party of the Eu seat of Merkel could stand" and in saying this he points both hands away from the bust towards his left, probably to make clear the geographical distance between Italy and Germany of Merkel. The third integrative gesture is instead used to make explicit a list of issues addressed: "In our country and in Europe 3 very profound things have happened. The longest economic crisis since the end of World War II. The most serious refugee and displaced persons crisis since the end of World War II. And, in Europe, the most serious security crisis since the end of World War II with the bombs

that exploded and made all of us Europeans cry, for the dead in the streets of our capitals" and in saying this with the left hand the gesture of the number three is made.

Angelino Alfano: C'è qualcosa di più grave e di più profondo di cui mi sono occupato da Ministro dell'Interno. Perché io ho gestito l'immigrazione e ho gestito l'ordine pubblico. [integrating] Ed io sono stato accusato ed una delle cose che ha segnato la mia carriera politica, è la gestione della vicenda immigrazione. Alla quale non ha retto neanche il partito della sede Eu della Merkel [integrating] perché è calato di consenso proprio per l'immigrazione. Nel nostro paese e in Europa sono accadute 3 cose molto profonde. La più lunga crisi economica dalla fine della Seconda Guerra Mondiale. La più grave crisi dei profughi e rifugiati dalla fine della Seconda Guerra Mondiale. E, in Europa, la più grave crisi di sicurezza dalla fine della Seconda Guerra Mondiale con le bombe che sono esplose e hanno fatto piangere tutti noi europei, per i morti nelle strade delle nostre capitali. [integrating]

As you can again see from the results shown in the table both *reinforcing* and *supplementary* type of gesture-speech relationship are little used.

Reinforcing type of gesture-speech relationship are mainly used to reiterate a concept already expressed linguistically, as in the case of Angelino Alfano who turns out to be the interviewee who makes most use of this type of gesture. In this example, Alfano, talking about the consensus obtained by one of his rivals Matteo Salvini, claims that this consensus was obtained at his expense. So, in saying "against me", the open hands are close to his bust to underline his person.

Angelino Alfano: Quindi la sfida di Salvini avendo aggregato consenso - contro di me peraltro [reinforcing] - sull'immigrazione e dopo averlo aggregato incanalarlo su un regime di legislazione democratica in modo tale che se sarà maggioranza, se avrà la possibilità, quelle pulsioni si scarichino dentro un canale democratico.

As mentioned above, supplementary gestures are also used with a very

low frequency, they typically adds new information not coded in the linguistic content.

For example - in the interview with Simone Di Stefano contained in the corpus - the interviewee is asked to clarify the alleged relations of the party with a convicted member of the Mafia. The interviewee tries to explain without arriving at a satisfactory answer for the interviewer who continues to press him. At this point the interviewee lowers his gaze and moves his open right hand away from his torso while saying "but I don't want to escape" as if to implicitly ask the journalist to stop his suppositions and let him explain his position.

*Simone Di Stefano: **Ma io non voglio evadere** [supplementary], mi ascolti, mi faccia dire.*

Complementary type of gesture-speech relationship, so the gesture that bring a necessary complement to the incomplete linguistic information provided by the verbal message, are instead used more frequently by the respondents in the corpus under analysis, in most cases to disambiguate the message or simply some linguistic elements.

For example, at the beginning of the interview with Carlo Calenda, he is shown a photo that portrays him wearing a worker's helmet; the interviewee refers to the photo by pointing with his left hand away from his torso to the screen where the photo appeared, making it easier for viewers to understand what he was referring to.

*Carlo Calenda: No no! Io intanto considero il fatto di essere considerato operaista, un grande complimento. Benché gli operai non si sentiranno come posso dire contenti dopo aver visto **la mia foto con quel caschetto** [complementary] in cui sembravo un totale ebete. Però detto questo, questo è un problema a parte.*

Instead it is curious to note that *contradictory* type of gestures has never been found in our dataset. The politicians interviewed never use gestures that contradict what they have said. Probably some contradiction can be found in their own words (as can be guessed from some sentences) but

such conclusions could be reached only after an in-depth analysis of the content and coherence of the interviews which, however, is not the object of this work.

As noted above, a residual category has been added to the tags. In the category *other* fall all those gestures that the annotator was not able to classify with the above mentioned semantic labels. As noted above, a residual category has been added to the tags .

This problem was found most frequently in the interview with Carlo Calenda, in fact, of the 103 hand movements noted, 26 fell into this category. Actually, it was possible to notice that these gestures assume a *ba-tonic* value, that is, in almost all cases they are used to mark the rhythm of enunciation by tapping a finger on the table, for example.

5.2.1 Political party influences the type of gestures: the one-way ANOVA test

As a result of these analyses, we sought to understand whether there was a significant relationship between the political party of affiliation and the type of gestures used, and thus whether the political party of affiliation influenced the use of one category of gestures rather than another. Before proceeding with the one-way ANOVA test (acronym for ANalysis Of VAriance) with independent samples, to make the political parties comparable we normalize the gestures manually annotated by the number of turns uttered by each politician as suggested by (Colletta et al., 2015) and made by (Lin, 2017).

Given the initial null hypothesis, i.e. that membership of a political party **does not** influence the type of gesture-speech relationship, the results obtained suggest that this hypothesis can be rejected since the significance values obtained $p (0,03)$ are lower than the conventional threshold value 0,05. Moreover, since the F value: 3,07 is higher than the F crit: 2,75,

Origin of variation	SQ	gdl	MQ	F	p-value	F crit.
Inter-groups	7167,66	4	1791,91	3,07	0,03	2,75
Intra-groups	14555,83	25	582,23			
Total	21723,5	29				

TABLE 5.6: ANOVA test results

we have another evidence in favor of the rejection of the null hypothesis. Therefore we can conclude that the party affiliation significantly influences the gesture-speech relationship.

Recall that in statistics, the F-test for comparing two variances is a hypothesis test based on the Fisher-Snedecor F-distribution and aimed at testing the hypothesis that two populations that both follow normal distributions have the same variance.

Chapter 6

Conclusions, future work and perspectives in computational linguistics

In this work, the creation of PoliModal corpus, the first freely-available multimodal corpus of political interviews, is described. The corpus - manually annotated with six non-verbal traits - covers 56 interviews, where each guest is associated with a role (for non-politicians) or a political party. Although a preliminary version of the resource was already presented in Trotta et al., 2019, several changes and extensions have been made. In particular, new features, including an additional annotation layer for the gesture, posture, and facial displays, are described.

The annotation scheme, inspired by the TEI-xml standard, is presented and described in detail. A first validation is made using the inter-annotator agreement. Furthermore, an experiment is proposed to test the robustness and reliability of the multimodal annotation schema.

A case study involving a sample of three participants having a different political orientation is presented.

The decision to focus on a small sample of analysis to study the correlation between particular semantic types of gestures and speech is motivated by the high cost in terms of time and computation involved in the

process of annotation and revision.

Although the corpus annotated so far does not allow for generalizations, we can already observe how the three politicians adopt different communication strategies, with Renzi being more emotional and showing more multimodal trails, while the other two are colder and Salvini tends to express his thoughts exclusively through lexical and semi lexical traits. This preliminary analysis shows the potential of putting different modalities in relation as a means to have a wider perspective on political discourse and persuasive strategies.

Subsequently, a set of statistical analyses of the traits and their association with language complexity and with the speakers' political orientation are presented. The results show that differences pertain more to single persons and conversational style than to political orientation.

These experiments led to the release of the annotation guidelines for annotating politics-domain spoken language. The resource is also released on principal repositories, including Github, Clarin, and Accademia della Crusca.

After that, the focus of the work shifts to co-gesture analysis. In order to demonstrate co-speech gestures of several Italian politicians during face-to-face interviews, labels describing the semantic type of the different hand movements were added to the corpus. Concerning gesture-speech relationships, hand movements turn out to be mainly used with integrative and complementary functions. So, the information provided by such gestures adds precision and emphasis to spoken information. In addition, results suggest that party affiliation does not significantly influence the gesture-speech relationship.

Finally, the Lexical Retrieval Hypothesis is tested by computing the association between hand movements produced by each interviewee and speech disfluencies using weighted mutual information. Results show that hand movements tend to co-occur with full pauses (i.e. repetition)

and empty pauses (i.e. pause) and more frequently with interjections (i.e. semi-lexical), suggesting that gesticulating may represent an attempt at lexical retrieval.

Future developments include mainly:

- the expansion of the resource in terms of interviews to be included and the expansion of semantic annotation to the entire existing corpus
- further analyses aimed at understanding whether such gestures co-occur with specific types of words (e.g. copulative verbs, predicative verbs, etc.) and whether other linguistic or socio-linguistic variables such as language complexity or age influence the use of hand movements and their semantic functions.

In conclusion, the application that such an annotated resource might have in the field of computational linguistics deserves a separate discussion.

6.0.1 Machine learning algorithms and multimodal corpora

Natural face-to-face communication - as extensively discussed above - is inherently multimodal. Humans continuously change their speech and body behaviors, adapting them to different communicative contexts, situations, and interlocutors. However, studies and applications using multimodal data are still very few in the literature.

This lack is even more evident by moving to languages other than English. The creation of resources is a critical problem that seriously plagues especially low-resource languages, such as Italian. In the case of multimodal resources, the problem is further complicated by the need to integrate and annotate information on different levels.

With the rise of unsupervised approaches, different techniques based on machine-learning algorithms have been successfully adapted to multimodal corpora. Notice that these approaches use transfer learning (Navarretta, 2013) to compensate for the lack of dedicated multimodal resources. The classifier is first trained on general-purpose datasets and then is used to annotate data in a new domain. It has proven to be a feasible way to reuse data and reduce annotation costs.

For instance, in (Jokinen and Ragni, 2007; Jokinen, Navarretta, and Paggio, 2008), machine learning algorithms have been used to recognize some of the functions of head movements in annotated corpora from various languages. Other studies have been focused on human-agent communication (Fujie et al., 2004; Morency et al., 2005; Morency et al., 2007; Morency, Kok, and Gratch, 2010). Other works have proposed techniques to improve classification performance in domain adaptation or transfer in various fields (Blitzer, McDonald, and Pereira, 2006; Moore and Lewis, 2010; Saenko et al., 2010).

Although these works have highlighted the potential of combining semi-automatic approaches to multimodal resources, cross-domain and transfer-learning approaches do not always guarantee adequate performance. In addition, they need very large datasets to achieve accurate performance. Finally, another open issue concerns the interpretability of the results obtained via unsupervised approaches. This weak point precludes the way for qualitative or more linguistics-oriented analyses.

In this context, the resource presented in this thesis work could contribute to training a baseline for the Italian language for machine learning algorithms. Furthermore, it can improve the performance of existing algorithms for specific tasks. For instance, current Argument Mining methodologies (Mancini et al., 2022) aiming at identifying argumentative structures in text or using persuasion techniques, particularly in the political context, might find great help in using an annotated resource. It becomes

even more valuable given the current scenario for the Italian language, in which similar studies or resources are almost entirely absent.

Bibliography

- Adolphs, Svenja and Ronald Carter (2013). *Spoken corpus linguistics: From monomodal to multimodal*. Routledge.
- Alfano, Iolanda et al. (2014). "VOLIP: a Corpus of Spoken Italian and a Virtuous Example of Reuse of Linguistic Resources". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA). ISBN: 978-2-9517408-8-4.
- Allwood, Jens (2001). "Dialog Coding-Function and Grammar: Göteborg Coding Schemsas". In: *rapport nr.: Gothenburg Papers in Theoretical Linguistics 85*.
- (2008). "Multimodal Corpora". In: *Corpus Linguistics. An International Handbook*. Ed. by Lüdeling, A. & Kytö, and M. Mouton de Gruyter, pp. 207–225. URL: <https://hal-hprints.archives-ouvertes.fr/hprints-00511882>.
- Allwood, Jens et al. (2007). "The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena". In: *Language Resources and Evaluation 41.3-4*, pp. 273–287.
- Austin, John Langshaw (1975). *How to do things with words*. Vol. 88. Oxford university press.
- Bartolini, Roberto et al. (2014). "From Synsets to Videos: Enriching Ital-WordNet Multimodally." In: *LREC*, pp. 3110–3117.
- Bazzanella, Carla (1992). "Aspetti pragmatici della ripetizione dialogica". In.

- Berretta, Monica (1994). "Il parlato italiano contemporaneo". In: *Storia della lingua italiana* 2, pp. 239–270.
- Biber, Douglas (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Bigi, Brigitte et al. (2011). "Multimodal annotations and categorization for political debates". In: *ICMI Workshop on Multimodal Corpora for Machine learning*, pp. 1–4.
- Blitzer, John, Ryan McDonald, and Fernando Pereira (2006). "Domain adaptation with structural correspondence learning". In: *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128.
- Bolasco, Sergio, Nora Galli de'Paratesi, and Luca Giuliano (2006). *Parole in libertà: un'analisi statistica e linguistica dei discorsi di Berlusconi*. Manifestolibri.
- Bongelli, Ramona, Ilaria Riccioni, and Andrzej Zuczkowski (2010). "Certain-uncertain, true-false, good-evil in Italian political speeches". In: *International Workshop on Political Speech*. Springer, pp. 164–180.
- Bressemer, Jana and Silva H Ladewig (2011). "Rethinking gesture phases: Articulatory features of gestural movement?" In: *Semiotica* 2011.184, pp. 53–91.
- Bruti, Silvia (2016). "Teaching compliments and insults in the EFL classroom through film clips". In: *Pragmatic issues in specialized communicative contexts*. Brill Rodopi, pp. 149–170.
- Butcher, Cynthia (2000). "two-word speech: when hand and mouth come together". In: *Language and gesture* 2, p. 235.
- Butterworth, Brian and Geoffrey Beattie (1978). "Gesture and silence as indicators of planning in speech". In: *Recent advances in the psychology of language*. Springer, pp. 347–360.
- Calzolari, Nicoletta et al. (2012). "The LRE Map. Harmonising Community Descriptions of Resources." In: *LREC*, pp. 1084–1089.

- Cassell, Justine et al. (2000). "Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents". In: *Embodied conversational agents* 1.
- Catellani, Patrizia, Mauro Bertolotti, and Venusia Covelli (2010). "Counterfactual communication in politics: Features and effects on voters". In: *International Workshop on Political Speech*. Springer, pp. 75–85.
- Cedroni, Lorella (2010). "Politolinguistics. Towards a New Analysis of Political Discourse". In: *International Workshop on Political Speech*. Springer, pp. 220–232.
- Chilton, Paul Anthony (2004). *Analysing political discourse: Theory and practice*. Psychology Press.
- Cichocka, Aleksandra et al. (2016). "On the grammar of politics—or why conservatives prefer nouns". In: *Political Psychology* 37.6, pp. 799–815.
- Cienki, Alan and Cornelia Müller (2008). *Metaphor and gesture*. Vol. 3. John Benjamins Publishing.
- Colletta, Jean-Marc et al. (2015). "Effects of age and language on co-speech gesture production: an investigation of French, American, and Italian children's narratives". In: *Journal of child language* 42.1, pp. 122–145.
- Coombs, Clyde Hamilton, Robyn M Dawes, and Amos Tversky (1970). "Mathematical psychology: An elementary introduction." In.
- Cresti, Emanuela and Alessandro Panunzi (2013). *Introduzione ai corpora dell'italiano*. Il mulino.
- Cronbach, Lee J (1955). "On the non-rational application of information measures in psychology". In: *Information Theory in Psychology Problems and Methods*; Quastler, H., Ed, pp. 14–30.
- Dinarelli, Marco et al. (Mar. 2009). "Annotating Spoken Dialogs: From Speech Segments to Dialog Acts and Frame Semantics". In: *Proceedings of SRSL 2009, the 2nd Workshop on Semantic Representation of Spoken Language*. Athens, Greece: Association for Computational Linguistics, pp. 34–41. URL: <https://www.aclweb.org/anthology/W09-0505>.

- Dittmann, Allen T and Lynn G Llewellyn (1969). "Body movement and speech rhythm in social conversation." In: *Journal of personality and social psychology* 11.2, p. 98.
- Dobrogaev, SM (1929). "Uchenie o refleksie v problemakh iazykovedeniia [Observations on reflexes and issues in language study]". In: *Iazykovedenie i materializm*, pp. 105–173.
- Ducrot, Oswald (1995). "Les modificateurs déréalisants". In: *Journal of pragmatics* 24.1-2, pp. 145–165.
- Duncan, S (2004). "Coding manual". In: *Disponível:< http://mcneilllab.uchicago.edu/pdfs/Coding_Manual.pdf>. Acesso em 28.*
- D'Agostino, Emilio, Annibale Elia, and Simonetta Vietri (2004). "Lexicon-grammar, electronic dictionaries and local grammars of italian". In: *Linguisticae Investigationes Supplementa* 24, pp. 125–136.
- D'Errico, Francesca, Isabella Poggi, and Laura Vincze (2010). "Discrediting body. A multimodal strategy to spoil the other's image". In: *International Workshop on Political Speech*. Springer, pp. 181–206.
- Ekman, Paul and Wallace V Friesen (1972). "Hand movements". In: *Journal of communication* 22.4, pp. 353–374.
- Esposito, Fabrizio et al. (2015). "The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus". In: *Proceedings of CLiC-it*.
- Fairclough, Norman (1998). "Political discourse in the media: An analytical framework". In: *Approaches to media discourse*, pp. 142–162.
- (2000). *New Labour, new language?* Psychology Press.
- Fujie, Shinya et al. (2004). "A conversation robot using head gesture recognition as para-linguistic information". In: *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*. IEEE, pp. 159–164.
- Furkó, Péter and Ágnes Abuczki (2014). "English discourse markers in mediatised political interviews". In.

- Givón, Talmy (1995). *Functionalism and grammar*. John Benjamins Publishing.
- Gross, Maurice (1994). *Constructing lexicon-grammars*. Centre national de la recherche scientifique, Universités de Paris 7 et 8.
- Guarasci, Raffaele et al. (2020). "Lexicon-Grammar based open information extraction from natural language sentences in Italian". In: *Expert Systems with Applications* 143, p. 112954.
- Guerini, Marco, Carlo Strapparava, and Oliviero Stock (2008). "Corps: A corpus of tagged political speeches for persuasive communication processing". In: *Journal of Information Technology & Politics* 5.1, pp. 19–32.
- Guiasu, Silviu (1977). *Information theory with applications*. McGraw-Hill Companies.
- Henderson, Alan, Frieda Goldman-Eisler, and Andrew Skarbek (1966). "Sequential temporal patterns in spontaneous speech". In: *Language and Speech* 9.4, pp. 207–216.
- Heritage, John (1985). "Analyzing news interviews: Aspects of the production of talk for an 'overhearing' audience". In: *Handbook of Discourse Analysis, vol. III: Discourse and Dialogue*.
- Hickmann, Maya (2002). *Children's discourse: person, space and time across languages*. Vol. 98. Cambridge University Press.
- Hostetter, Autumn B, Martha W Alibali, and Sotaro Kita (2007). "I see it in my hands' eye: Representational gestures reflect conceptual demands". In: *Language and Cognitive Processes* 22.3, pp. 313–336.
- Jaiswal, Mimansa et al. (2020). "Muse: a multimodal dataset of stressed emotion". In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1499–1510.
- Jansen, Michel-Pierre et al. (2020). "Introducing MULAI: A multimodal database of laughter during dyadic interactions". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4333–4342.

- Jezek, Elisabetta et al. (2014). "T-PAS: A resource of corpus-derived Types Predicate-Argument Structures for linguistic analysis and semantic processing". In: *Proceedings of LREC*, pp. 890–895.
- Jokinen, Kristiina (2020). "The AICO Multimodal Corpus–Data Collection and Preliminary Analyses". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 559–564.
- Jokinen, Kristiina, Costanza Navarretta, and Patrizia Paggio (2008). "Distinguishing the Communicative Functions of Gestures: An Experiment with Annotated Gesture Data". In: *Machine Learning for Multimodal Interaction: 5th International Workshop, MLMI 2008, Utrecht, The Netherlands, September 8-10, 2008. Proceedings 5*. Springer, pp. 38–49.
- Jokinen, Kristiina and Anton Ragni (2007). "Clustering experiments on the communicative properties of gaze and gestures". In: *Proceeding of the 3rd. Baltic Conference on Human Language Technologies*.
- Kendon, Adam (1972). "Some relationships between body motion and speech". In: *Studies in dyadic communication* 7.177, p. 90.
- (1980). "Gesticulation and speech: Two aspects of the". In: *The relationship of verbal and nonverbal communication* 25, p. 207.
- (2004). *Gesture: Visible action as utterance*. Cambridge University Press.
- Kipp, Michael (2001). "Anvil—a generic annotation tool for multimodal dialogue". In: *Seventh European Conference on Speech Communication and Technology*.
- Knight, Dawn (2011). *Multimodality and active listenership: A corpus approach*. A&C Black.
- Kontogiorgos, Dimosthenis, Elena Sibirtseva, and Joakim Gustafson (2020). "Chinese whispers: A multimodal dataset for embodied language grounding". In: *Language Resources and Evaluation (LREC)*.
- Koutsombogera, Maria and Harris Papageorgiou (2010). "Multimodal indicators of persuasion in political interviews". In: *International Workshop on Political Speech*. Springer, pp. 16–29.

- Krauss, Robert M (1998). "Why do we gesture when we speak?" In: *Current directions in psychological science* 7.2, pp. 54–54.
- Krauss, Robert M and Uri Hadar (1999). "The role of speech-related arm/hand gestures in word retrieval". In: *Gesture, speech, and sign* 93.
- Krippendorff, Klaus (2018). *Content analysis: An introduction to its methodology*. Sage publications.
- Kvålseth, Tarald O (1991). "The relative useful information measure: Some comments". In: *Information sciences* 56.1-3, pp. 35–38.
- Laustsen, Lasse and Michael Bang Petersen (2016). "Winning faces vary by ideology: How nonverbal source cues influence election and communication success in politics". In: *Political Communication* 33.2, pp. 188–211.
- Lenci, Alessandro, Gabriella Lapesa, and Giulia Bonansinga (2012). "LexIt: A Computational Resource on Italian Argument Structure." In: *LREC*, pp. 3712–3718.
- Levelt, Willem JM (1983). "Monitoring and self-repair in speech". In: *Cognition* 14.1, pp. 41–104.
- Lin, Yen-Liang (2017). "Co-occurrence of speech and gestures: A multimodal corpus linguistic approach to intercultural interaction". In: *Journal of Pragmatics* 117, pp. 155–167.
- Lindsay, James R (1975). "Producing simple utterances: How far ahead do we plan?" In: *Cognitive Psychology* 7.1, pp. 1–19.
- Lockhead, GR (1970). "Identification and the form of multidimensional discrimination space." In: *Journal of Experimental Psychology* 85.1, p. 1.
- Longobardi, Ferdinando (2010). "Linguistic Factors in Political Speech". In: *International Workshop on Political Speech*. Springer, pp. 233–244.
- Lucisano, Pietro and Maria Emanuela Piemontese (1988). "GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana". In: *Scuola e città* 3.31, pp. 110–124.

- Mancini, Eleonora et al. (2022). "Multimodal Argument Mining: A Case Study in Political Debates". In: *Proceedings of the 9th Workshop on Argument Mining*, pp. 158–170.
- Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier (2015). "The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment". In: *Computational Linguistics* 41.3, pp. 437–479.
- McNeill, D (2005). *Gesture and thought*. Chicago, IL, US.
- McNeill, David (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- Menini, Stefano et al. (2020). "DaDoEval@ EVALITA 2020: Same-genre and cross-genre dating of historical documents". In: *7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. EVALITA 2020*. Accademia University Press, pp. 391–397.
- Miller, James Edward, Jim Miller, Regina Weinert, et al. (1998). *Spontaneous spoken language: Syntax and discourse*. Oxford University Press on Demand.
- Moneglia, Massimo et al. (2014). "The IMAGACT visual ontology. An extendable multilingual infrastructure for the representation of lexical encoding of action". In.
- Moore, Robert C and William Lewis (2010). "Intelligent selection of language model training data". In: *Proceedings of the ACL 2010 conference short papers*, pp. 220–224.
- Morency, Louis-Philippe, Iwan de Kok, and Jonathan Gratch (2010). "A probabilistic multimodal approach for predicting listener backchannels". In: *Autonomous agents and multi-agent systems* 20, pp. 70–84.
- Morency, Louis-Philippe et al. (2005). "Contextual recognition of head gestures". In: *Proceedings of the 7th international conference on Multimodal interfaces*, pp. 18–24.

- (2007). “Head gestures for perceptual interfaces: The role of context in improving recognition”. In: *Artificial Intelligence* 171.8-9, pp. 568–585.
- Morsella, Ezequiel and Robert M Krauss (2004). “The role of gestures in spatial working memory and speech”. In: *The American journal of psychology*, pp. 411–424.
- Nakamura, Kai, Sharon Levy, and William Yang Wang (2019). “r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection”. In: *arXiv preprint arXiv:1911.03854*.
- Navarretta, Costanza (2013). “Transfer learning in multimodal corpora”. In: *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, pp. 195–200.
- Navarretta, Costanza and Patrizia Paggio (2010). “Multimodal Behaviour and Interlocutor Identification in Political Debates”. In: *International Workshop on Political Speech*. Springer, pp. 99–113.
- Paetzel, Maike, Deepthi Karkada, and Ramesh Manuvinakurike (2020). “Rdg-map: A multimodal corpus of pedagogical human-agent spoken interactions”. In: *12th Conference on Language Resources and Evaluation (LREC 2020), 11–16 May 2020, Marseille, France*, pp. 600–609.
- Palmero Aprosio, Alessio and Giovanni Moretti (2018). “Tint 2.0: an All-inclusive Suite for NLP in Italian”. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.
- Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi (2002). “Multi-WordNet: developing an aligned multilingual database”. In: *First international conference on global WordNet*, pp. 293–302.
- Pocock, Adam Craig (2012). “Feature selection via joint likelihood”. PhD thesis. University of Manchester.
- Poggi, Isabella (2005). “The goals of persuasion”. In: *Pragmatics & Cognition* 13.2, pp. 297–335.
- (2006). *Le parole del corpo: introduzione alla comunicazione multimodale*. Carocci editore.

- Portelli, Alessandro (1985). *Biografia di una città: storia e racconto: Terni 1830-1985*. Einaudi.
- Purpura, Stephen and Dustin Hillard (2006). "Automated classification of congressional legislation". In: *Proceedings of the 2006 international conference on Digital government research*. Digital Government Society of North America, pp. 219–225.
- Saenko, Kate et al. (2010). "Adapting visual category models to new domains". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, pp. 213–226.
- Salvati, Luisa and Massimo Pettorino (2010). "A Diachronic Analysis of Face-to-Face Discussions: Berlusconi, Fifteen Years Later". In: *International Workshop on Political Speech*. Springer, pp. 65–74.
- Schegloff, Emanuel A (2000). "Overlapping talk and the organization of turn-taking for conversation". In: *Language in society* 29.1, pp. 1–63.
- Schoonvelde, Martijn et al. (2019). "Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches". In: *PloS one* 14.2, e0208450.
- Seiter, John S and Harry Weger Jr (2020). *Nonverbal communication in political debates*. Lexington Books.
- Simone, Raffaele (1990). *Fondamenti di linguistica*. Laterza Bari.
- Sprugnoli, Rachele et al. (2016). "Fifty years of European history through the Lens of Computational Linguistics: the De Gasperi Project". In: *IJCOL*, pp. 89–100. URL: http://www.ai-ic.it/IJCoL/v2n2/5-sprugnoli_et_al.pdf.
- Stam, Gale and S McCafferty (2008). "Gesture studies and second language acquisition: A review". In: *Gesture: second language acquisition and classroom research*, pp. 3–24.
- Tannen, Deborah (1989). "Interpreting interruption in conversation". In: *Gender and discourse*, pp. 53–83.

- Tonelli, Sara, Rachele Sprugnoli, and Giovanni Moretti (2019). "Prendo la Parola in Questo Consesso Mondiale: A Multi-Genre 20th Century Corpus in the Political Domain". In: *Proceedings of CLIC-it*.
- Trotta, Daniela et al. (2019). "Annotation and Analysis of the PoliModal Corpus of Political Interviews". In: *Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*.
- Trotta, Daniela et al. (2020). "Adding Gesture, Posture and Facial Displays to the PoliModal Corpus of Political Interviews". In: *12th Language Resources and Evaluation Conference (LREC 2020)*. European Language Resources Association, pp. 4320–4326.
- Truong, Khiet P. (2013). "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee". In: *Proceedings of INTERSPEECH*.
- Tsui, Amy BM (1994). *English conversation*. Oxford University Press.
- Tuite, Kevin (1993). "The production of gesture". In: *SEMIOTICA-LA HAYE THEN BERLIN- 93*, pp. 83–83.
- Vignozzi, Gianmarco (2019). "HOW GESTURES CONTRIBUTE TO THE MEANINGS OF IDIOMATIC EXPRESSIONS AND PHRASAL VERBS IN TV BROADCAST INTERVIEWS: A multimodal analysis." In: *Lingue e Linguaggi* 29.
- Voghera, Miriam (1992). *Sintassi e intonazione nell'italiano parlato*. Il mulino.
- (2001). *Teorie linguistiche e dati di parlato*. na.
- Wagner, Petra, Zofia Malisz, and Stefan Kopp (2014). *Gesture and speech in interaction: An overview*.
- Yamazaki, Yoshihiro et al. (2020). "Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness". In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 443–448.

- Yoshioka, K (2008). "Linguistic and gestural introduction of inanimate referents in L1 and L2 narrative". In: *ESL & applied linguistics professional series*, pp. 211–230.
- Zurloni, Valentino and Luigi Anolli (2010). "Fallacies as argumentative devices in political debates". In: *International Workshop on Political Speech*. Springer, pp. 245–257.

