# Building a Pragmatically Annotated Diachronic Corpus: the DIADIta Project

Irene De Felice[1,*] and Francesca Strik-Lievers[2]

[1] *Università del Piemonte Orientale, Via Galileo Ferraris 116, Vercelli, Italy*
[2] *Università di Genova, Piazza Santa Sabina 1, Genova, Italy*

## Abstract

We present here the first stages of the construction of the DIADIta corpus, a diachronic corpus of Italian annotated for interactional pragmatic phenomena. This corpus aims to fill a gap in the resources available for the historical pragmatics of Italian. First, we describe the annotation scheme, which is structured into four levels covering a wide range of pragmatic (or pragmatically relevant) categories: speech acts (e.g., apology; threat), forms (e.g., discourse marker; expressive), pragmatic functions (which are speaker-oriented, e.g., mitigation; turn-taking), and pragmatic aims (which are interlocutor-oriented, e.g., attention-getting; request for agreement). We then discuss how the results of an initial annotation exercise provide insights for refining the annotation procedure.

## Keywords

diachronic corpus pragmatics, historical pragmatics, interaction, Italian, pragmatic annotation

## 1. Introduction

The DIADIta project[1], situated within the framework of historical pragmatics [1], aims to investigate the specific pragmatic features and strategies of dialogic interaction in different phases of the Italian language, and to understand how these features and strategies interrelate with one another and change over time. Although the last fifteen years have witnessed a growing interest in the historical pragmatics of Italian [2], there is still a lack of an in-depth study on this topic, one that is able to fully account for how different communicative strategies and different linguistic categories (primarily, but not exclusively, pragmatic) interact with each other, both in synchronic and diachronic perspective. The DIADIta project aims to address this gap.

A key goal of the project is to build a diachronic corpus annotated for a wide range of pragmatically relevant linguistic phenomena. The DIADIta corpus, which will contribute to the recently established field of diachronic corpus pragmatics [3], will consist of at least 24 Italian literary texts of different genres dating from the 13th to the 20th century: in most cases, plays, novels and short stories where dialogic interactions between characters are particularly frequent. Once completed, the corpus will be freely accessible and searchable from the project website (www.diadita.it) and will be possibly further expanded and enriched with other texts of different literary genres.

In this paper, we present the first steps we have taken to lay the foundation for the DIADIta corpus. After a brief review of related literature and resources (Section 2), we describe the structure of the annotation scheme, outlining the theoretical and methodological assumptions that underlie it and highlighting its most innovative aspects (Section 3). Then, we present the results of an annotation exercise on a play by Luigi Pirandello, with which we tested the reliability of the scheme. In the light of these results, we also briefly discuss some improvements that we plan to apply in the next stages of the corpus annotation process (Section 4). The last section draws the conclusions of the study (Section 5).

*CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy.*
* Corresponding author.
irene.defelice@uniupo.it (I. De Felice);
francesca.striklievers@unige.it (F. Strik-Lievers)


[1] PRIN 2022 project *Dialogic interaction in diachrony: a pragmatic history of the Italian language - DIADIta* (2023-2025), national P.I. Maria Napoli (Università del Piemonte Orientale), P.I. for the University of Genova Chiara Fedriani. The paper was conceived by the two authors together. For academic reasons only, the scientific responsibility is attributed as follows: Sections 2, 3.2, 3.3, 4 to Irene De Felice; Sections 1, 3, 3.1, 5 to Francesca Strik-Lievers.

## 2. Pragmatically annotated (diachronic) corpora: challenges and resources

Most existing corpora are not well suited for research focused on pragmatics, unless one adopts a form-to-function approach, which implies searching for specific keywords or linguistic structures that are known or supposed to express pragmatic functions (e.g. discourse markers, specific verb forms and syntactic structures, etc.; see [4, 5]). Such an approach is not viable in the field of diachronic pragmatics: in this case, a function-to-form approach must usually be adopted, since certain pragmatic functions remain stable over time, while the linguistic means by which speakers express them may vary [6, 7]. The problem is, of course, that "functions cannot be searched for automatically" [8, p. 5].

Corpora annotated with pragmatic information that allow for searches based on a function-to-form approach are rare, partly due to the difficulties arising in their construction [9]. First of all, the annotation of pragmatic categories requires a great deal of interpretation on the part of the annotator. Moreover, this type of annotation, "unlike, for example, POS (part-of-speech) or semantic tagging/annotation, almost always needs to take into account levels above the individual word and may even need to refer to contextual information beyond those textual units that are commonly referred to as a 'sentence' or 'utterance'" [10, p. 84]. Therefore, due to its inherent difficulties, the annotation of pragmatic categories is still mostly a manual, time-consuming task and "it is doubtful whether the process of manual classification will ever be fully replaced" [8, p. 15]. Nevertheless, some attempts have been made to design annotation schemes that allow for (semi-)automatic annotation of specific pragmatic categories. In particular, most efforts have focused on speech acts. Consider, for instance, the *Speech Act Annotated Corpus* project (SPAAC; [11]) and the *Dialogue Annotation and Research Tool* (DART; [12, 13]; for a discussion of widely known models and tools for speech act or dialogue act annotation, including the DAMSL and the SWBD-DAMSL models, see [10, 14], and more recently [15]). The international standard DiAML (*Dialogue Act Markup Language*, ISO 24617-2; see [16]) also concerns speech acts found in dialogue. In this annotation scheme, a given dialogue segment may express multiple acts, and a given act may be assigned multiple communicative functions: a feature that is also crucial in our annotation scheme (see Section 3.1).

Corpora annotated with pragmatic categories for English include, among others, the *SPICE-Ireland Corpus*, which is derived from the spoken data of the *International Corpus of English: Ireland Component* (*ICE-Ireland*) and provides information on the speech act function of utterances, discourse markers, and quotatives. The *Sociopragmatic Corpus* (SPC) is a subsection of the *Corpus of English Dialogues* (CED) and comprises drama and trial proceedings dating from 1640 to 1760. This historical corpus can be used to investigate the extent to which the role of the participants affects the realization of pragmatic functions [8], since gender, status/social rank, role, and age are annotated for each participant.

For the Italian language, there are numerous corpora that collect texts from historical varieties of Italian (e.g. DiaCORIS – *Corpus of Diachronic Written Italian*; CEOD – *Digital Nineteenth-Century Epistolary Corpus*), some of which also provide morphological information (e.g. MIDIA – *Morphology of Italian in Diachrony*). There are also corpora designed to enable or facilitate pragmatic analysis. For example, the LABLITA corpus [17], developed within the pragmatic framework of the *Language into Act Theory* (L-AcT), brings together in a single resource a collection of three spoken Italian corpora recorded in Tuscany since 1965. One of the most innovative aspects of the corpus is that the transcripts are aligned with the acoustic source via utterance, i.e., "the linguistic counterpart of a speech act" [17, p. 93]. Linguistic implicatures (presuppositions, implicatures, topicalizations, and vagueness) are annotated in the IMPAQTS corpus, which collects Italian political discourses since 1946 [18].

Although this is a brief and non-exhaustive overview of the resources in this field, the few examples provided are sufficient to demonstrate that, overall, it is still true what Archer and colleagues wrote in 2008, that is, that "[w]ork in the area of pragmatics and corpus annotation is much less advanced than other annotation work (grammatical annotation schemes, for example)" [19, p. 613]. Furthermore, to the best of our knowledge, a diachronic corpus annotated with a rich set of pragmatic features is currently lacking among the corpora developed for Italian, and we find no equivalents among the corpora developed for other languages either. Most notably, there is no resource capable of accounting for both the linguistic means that express different pragmatic functions in various historical varieties of a language, and the ways in which these linguistic categories interact with one another in both a synchronic and diachronic dimension. This led to the design and construction of the DIADIta corpus.

## 3. Annotation scheme

The annotation scheme created within the DIADIta project is designed to cover a wide range of pragmatically relevant phenomena, especially those with a clear interactional value. Given that no existing tagset fully met the project's needs to encompass a broad spectrum of linguistic—and particularly pragmatic—

phenomena, the annotation scheme has been developed by drawing from a number of categories whose relevance is well established in pragmatic studies, such as POLITENESS, DISCOURSE MARKERS, further enriched with other linguistic categories that proved to have significant implications on the pragmatic front, such as EPISTEMICITY and EVIDENTIALITY.

So far, the scheme is organized into four levels of annotation (for a detailed description of the individual tags, please refer to the DIADIta annotation guidelines available on the project's website):

- **Forms**: This level includes linguistic expressions (belonging to different parts of speech, and with variable extension) that have an interactional pragmatic value, and in particular: DISCOURSE MARKERS (e.g., *Senti, io me ne vado*, 'Listen, I'm leaving'), EXPRESSIVES (e.g., *Smettila, idiota!*, 'Stop it, you idiot!') and REPETITION, when it has a pragmatic value (e.g., *Lo giuro, lo giuro!*, 'I swear it, I swear it!', where the repetition intensifies the oath).
- **Pragmatic functions**: This level includes a set of categories that have (also, or exclusively) a pragmatic value, such as: POLITENESS, VAGUENESS, DISAGREEMENT, IMPOLITENESS, INTENSIFICATION, EPISTEMICITY, TURN-TAKING.
- **Pragmatic aims**: This level focuses on the reaction that the speaker intends to provoke in the interlocutors, for example attracting their attention (ATTENTION GETTING) or requesting their confirmation or manifestation of agreement (REQUEST FOR CONFIRMATION/AGREEMENT)[2].
- **Speech acts**: This level includes the main types of expressive (e.g., DERISION, PROTEST), directive (e.g., ORDER, REQUEST), commissive (e.g., COMMITMENT/PROMISE, THREAT), and assertive (e.g., ASSERTION, CORRECTION) speech acts.

Each of the four levels includes several tags (N=57), as summarized in Appendix A.

## 3.1. Interaction between categories

As illustrated by examples from Luigi Pirandello's play *Enrico IV* (1921), the same string of text can be annotated with multiple tags, either from the same level (ex. 1) or from a different level (ex. 2). Furthermore, a string of text tagged with a certain tag can contain a smaller string tagged with a different tag, either from the same level (ex. 3) or from a different level (ex. 4):

1. Di Nolli: *Lasciamo andare, lasciamo andare,* ***vi prego****.*
   Di Nolli: 'Let it go, let it go, I beg you.'
2. D. Matilde: […] *Non ti vedi in me, tu, là?*
   Frida: ***Mah!*** *Io, veramente…*
   D. Matilde: '[...] Don't you see yourself in me, there? '
   Frida: 'Well! I, actually…'
3. Bertoldo: […] *Ho detto bene: non era vestiario, questo, del mille e cinquecento*!
   Arialdo: ***Ma che mille e cinquecento!***
   Bertoldo: '[...] I said it right: this wasn't clothing from the fifteen hundreds!'
   Arialdo: 'What fifteen hundreds!'
4. Bertoldo (arrabbiandosi): ***Ma*** *me lo potevano dire,* ***per Dio santo****, che si trattava di quello di Germania e non d'Enrico IV di Francia!*
   Bertoldo (getting angry): 'But they could have told me, for God's sake, that it was about the one from Germany and not Henry IV of France!'

In ex. 1, *vi prego* 'I beg you' is labeled with two tags from the pragmatic functions level: it has both a POLITENESS function and an INTENSIFICATION function (it intensifies the force of the directive act expressed by the whole utterance).

In ex. 2, *Mah!* 'Well!' is tagged as a DISCOURSE MARKER (forms level) but is also considered an expression of EPISTEMICITY and DISAGREEMENT (functions level). By using this interjection, the character Frida expresses a low degree of certainty regarding the truth of Donna Matilde's statement, thus also demonstrating that she does not fully agree with her.

In ex. 3, the entire utterance by Arialdo, who mocks Bertoldo in front of his friends (speech act of DERISION), is labeled at the level of pragmatic functions as a manifestation of DISAGREEMENT and IMPOLITENESS. However, it also contains the DISCOURSE MARKER *ma che* 'what,' which is also labeled – again at the pragmatic functions level - as a TURN-TAKING marker.

In ex. 4, the whole utterance by Bertoldo is labeled as a PROTEST (speech acts level). Within this utterance, *ma* 'but' is labeled as a DISCOURSE MARKER (forms level) and as a TURN-TAKING marker (pragmatic functions level), and *per Dio Santo* 'for God's sake' is labeled with the tags EXPRESSIVE (forms level) and INTENSIFICATION

[2] To avoid overburdening the tagset, we have chosen to merge certain categories that, despite being well-defined on a theoretical level, are often difficult to distinguish in practice from other closely related functional categories, such as REQUEST FOR CONFIRMATION and REQUEST FOR AGREEMENT.

(pragmatic functions level), since it is used to strengthen the illocutionary force of the act itself.

### 3.2. Annotation tool

As shown in Section 3.1, allowing overlapping annotations from the same and different levels is essential to capture the multifunctionality of pragmatically relevant expressions and the interaction between linguistic and pragmatic categories. For instance, *Mah!* 'Well!' serves as a DISCOURSE MARKER that expresses DISAGREEMENT while also conveying EPISTEMICITY, in ex. 2 discussed above. Moreover, having multiple annotators work on the same text is necessary for identifying and discussing cases of disagreement, especially in the early stages of the project.

For collaborative projects of this type, a web-based tool is the most suitable instrument [20]. For this first annotation exercise, we chose INCEpTION [21], which allows the creation and easy modification of a tagset (in our case multiple tagsets, one for each annotation level) and the overlapping and nesting of different tags. The annotation performed on INCEpTION is of the standoff type: the texts are therefore not modified, and the annotations are stored in a separate document (see Finlayson & Erjavec [22, p. 178], who consider standoff annotation a best practice, compared to inline annotation).

As an example, Figure 1 presents a screenshot of an annotation, again on the play *Enrico IV*. A stratification of annotations can be observed, with the entire utterance *Senti: io non ho mai capito perché si laureino in medicina!* ('Listen: I have never understood why they graduate in medicine!') labeled as an EXCLAMATION speech act, *senti* 'listen' as a DISCOURSE MARKER with the pragmatic function of TURN-TAKING and the pragmatic aim of ATTENTION-GETTING.

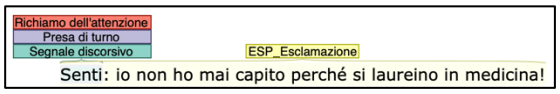


**Figure 1:** Screenshot of the annotation in INCEpTION. The four different colors represent different annotation layers: forms, pragmatic functions, pragmatic aims, speech acts.

### 3.3. Annotation guidelines

As the annotation scheme and the few examples provided in Section 3.1 clearly demonstrate, the annotation of the DIADIta corpus is extremely complex. Indeed, Weisser [10, p. 84] observes, "[a]ny type of linguistic annotation is a highly complex and interpretive process, but none more so than pragmatic annotation". Therefore, it is essential to have a meticulously detailed annotation manual to guide annotators.

The first text tested for the pragmatic annotation of the categories initially selected for our project is the first act of Pirandello's *Enrico IV* (9,216 words). We began by independently annotating the text and subsequently discussed our work until a consensus was reached on each annotation.

The total number of annotations for the first act is 958. This very first phase of the annotation process has been crucial for refining the tagset, which is now in the form shown in Appendix A, and for developing guidelines with practical instructions for annotation. The current version of the DIADIta annotation guidelines is available on the project's website. The guidelines provide a brief definition for each annotation level and tag, along with basic references and examples from the annotated texts in the corpus. They also specify constraints for applying certain tags. For example, the tag EXPRESSIVE (forms level) is used to annotate lexical elements such as exclamations, vulgarisms, insults, or curses that express "subjective sensations, emotions, affections, evaluations or attitudes" [23, p. 33]. However, it is also specified that this tag should only be applied when it co-occurs with one or more tags from the pragmatic functions or pragmatic aims levels; i.e, only in contexts where expressive forms are relevant at a pragmatic, interactional level. Consider examples 5 and 6:

5. Secondo valletto: *Eh,* ***santo Dio****, potevate dircelo!*
   Second valet: 'Oh, holy God, you could have told us!'
6. Frida: *Fa di professione lo* ***scemo****, non lo sa?*
   Frida: 'He acts the fool professionally, don't you know?'

In ex. 5, *santo Dio* 'holy God' is tagged as EXPRESSIVE because it also has an INTENSIFICATION function, as it intensifies the expressive force of a PROTEST speech act. In contrast, in ex. 6, *scemo* 'fool', despite being an expressive used in a DERISION speech act, is not tagged because it does not seem to serve primarily a specific pragmatic function or aim in the interaction.

## 4. Results and discussion

To test the reliability of the adopted scheme, we annotated the second act of Pirandello's *Enrico IV* (6,968 tokens) in INCEpTION. This annotation process benefited from our previous joint annotation experience on the first act of the same play and, most importantly, relied on the established annotation guidelines. The annotation performed separately by the two authors

resulted in 818 and 906 annotations, respectively, for a total of 1,724 annotations.

To test the inter-annotator agreement we adopted Krippendorff's α metric [24, 25, 26, 27], a unitizing measure that is particularly suitable for assessing the level of agreement in our case, because it can produce partial agreement scores from all annotations by also taking into account their partial overlaps. For instance, for *eh sì* ('oh, yes'), one annotator assigned the tag AGREEMENT (pragmatic functions) to the entire expression, while the other annotator assigned the same tag only to *sì*. This kind of annotation is considered incomplete, but is still used to compute the agreement. The agreement score is, of course, lower in such cases compared to complete annotations, where the same tag is assigned to the same length of spans by both annotators. Table 1 presents the agreement scores and the number of annotations for each of the four layers of our annotation scheme[3].

**Table 1**
Number of annotations and IAA scores (Krippendorff's α; α value may range from -1 to 1). FSL=Francesca Strik-Lievers, IDF=Irene De Felice.

| | FSL | IDF | Krippendorff's α |
|---|---|---|---|
| Forms | 171 | 168 | 0.71 |
| Functions | 327 | 390 | 0.34 |
| Aims | 29 | 38 | 0.05 |
| Speech acts | 291 | 310 | 0.56 |

According to Landis and Koch's [28] scale, our levels of agreement should be considered as *slight* for the pragmatic aims level, *fair* for the functions level, *moderate* for the speech acts level, and *substantial* for the forms level.

These results clearly demonstrate that, even though the annotation was performed by expert annotators following detailed guidelines, pragmatic annotation remains a highly complex and fine-grained task, especially when annotators have to assign many labels, and often multiple labels to the same token(s). In many cases, to understand the pragmatic function of a linguistic unit, the annotator must go well beyond the level of the single word, phrase or sentence, and necessarily consider the linguistic co-text, or even the extralinguistic context, as far as it can be reconstructed from a written text. Therefore, in this specific field of annotation, reaching an α value higher than 0.67, which is sometimes considered essential to draw at least "tentative conclusions" [24, p. 241] in other computational linguistic tasks, may be exceptionally challenging, even for expert annotators. Other complex pragmatic annotation models created for discourse annotation tasks have also failed to achieve high levels of agreement. For instance, *slight* to *moderate* values of agreement produced by the α metric are also reported by Duran et al. [27] for the *Conversation Analysis Modeling Schema - CAMS* (cf. also Castagneto [14], who reports moderate agreement values for the Chiba and DAMSL annotation models).

Therefore, a low level of agreement was to be expected and, from our point of view, this should not necessarily be understood as an indication of low annotation quality, inadequate training, or poorly defined guidelines [29], since when there are two partially or completely disagreeing annotations, it is not always the case that one is correct and the other wrong. In many cases both can be acceptable, as in example 7, in which Matilde's reaction to the doctor's question was considered by one annotator as an EXCLAMATION, and by the other as a RESPONSE to his request for information:

7. Dottore (stordito): *Come dice?*
   D. Matilde: ***Quest'automobile, dottore!*** *Sono più di tre ore e mezzo!*
   Doctor (stunned): 'What did you say?'
   D. Matilde: 'This car, doctor! It's been over three and a half hours!'

Discrepancies may also stem from differences in annotated span lengths, even when the same tag is chosen. For instance, in example 8, one annotator marked AGREEMENT for the entire statement by Belcredi (*Sì, forse, quando disse…*), while the other one marked AGREEMENT only for *sì* 'yes'.

8. D. Matilde: *Non è vero! – Di me! Parlava di me!*
   Belcredi: ***Sì, forse, quando disse****…*
   D. Matilde: *Dei miei capelli tinti!*
   D. Matile: 'That's not true! Me! He was talking about me!'
   Belcredi: 'Yes, maybe, when he said…'
   D. Matilde: 'About my dyed hair!'

The analysis of cases of disagreement has been also useful in order to revise certain aspects of the tagset. For instance, after this exercise we have decided to merge the COMMITMENT/PROMISE speech act with OATH in future annotations, given that in many cases it is very difficult to distinguish between them. It has also been useful to identify unclear points in the guidelines, and to better plan the next phases of the project. In particular, we intend to: (i) release an updated version of the guidelines with clearer descriptions of some aspects of the

[3] The inter-annotator agreement is calculated with INCEpTION 33.3-SNAPSHOT (b5644aca).

annotation process; (ii) ensure that each text in the corpus is annotated or revised by at least two expert annotators; and (iii) include validation tasks at a regular rate in the project workflow to revise annotations for small groups of texts in order to reach better intra- and inter-text consistency.

## 5. Conclusions

This paper has outlined the initial steps in creating the DIADIta corpus, a pragmatically annotated diachronic corpus for Italian. This corpus is characterized by its rich, multi-layered annotation scheme organized into four dimensions: forms, pragmatic functions, pragmatic aims, speech acts. This structure allows for nuanced analysis of pragmatic strategies in literary texts from the 13th to the 20th century. The innovative approach of annotating complex interactional features highlights the value of this corpus as an unparalleled tool for examining the evolution of pragmatic functions and forms over time, enabling detailed and multi-dimensional analysis of text data.

We have also detailed an annotation exercise on a play by Pirandello that illustrates the task's complexity (reflected in the low level of agreement in some layers), but also the richness of the annotations. This first exercise is crucial for refining the annotation process and improving clarity and reliability in applying a pragmatic annotation model to historical texts.

## Acknowledgements

Funded by the European Union - Next Generation EU (Mission 4, Component 1, CUP D53D23009600006) within the PRIN 2022 project *Dialogic interaction in diachrony: a pragmatic history of the Italian language – DIADIta* (2023-2025; P.I. Maria Napoli, Università del Piemonte Orientale). We thank Maria Napoli and Chiara Fedriani for their useful suggestions, together with the other members of our research group, Luisa Brucale, Ludovica Maconi and Giada Parodi, for the valuable moments of constructive discussion we have had. We also owe a special thanks to Richard Eckart de Castilho for his generous assistance with INCEpTION.

## References

[1] A. H. Jucker (Ed.), Historical Pragmatics. Pragmatic Developments in the History of English, John Benjamins, Amsterdam/Philadelphia, 1995.

[2] G. Alfieri, G. Alfonzetti, D. Motta, R. Sardo (Eds.), Pragmatica storica dell'italiano. Modelli e usi comunicativi del passato, Cesati, Firenze, 2020.

[3] I. Taavitsainen, A. H. Jucker, J. Tuominen (Eds.), Diachronic corpus pragmatics, John Benjamins, Amsterdam, 2014.

[4] J. Culpeper, M. Kytö, Early Modern English dialogues: Spoken interaction as writing. Studies in English Language, Cambridge University Press, Cambridge, 2010.

[5] U. Lutzky, Discourse markers in Early Modern English, John Benjamins, Amsterdam, 2012.

[6] A. H. Jucker, History of English and English Historical Linguistics, Ernst Klett, Stuttgart, 2000.

[7] A. H. Jucker, Corpus pragmatics, in: J.-O. Östman, J. Verschueren (Eds.), Handbook of Pragmatics, Benjamins, Amsterdam/Philadelphia, 2013, pp. 1–17.

[8] D. Landert, D. Dayter, T. C. Messerli, M. A. Locher, Corpus Pragmatics, Cambridge University Press, Cambridge, 2023.

[9] C. Rühlemann, What can a corpus tell us about pragmatics, in: A. O'Keeffe, M. J. McCarthy (Eds.), The Routledge Handbook of Corpus Linguistics, Routledge, New York, 2022, pp. 263–280.

[10] M. Weisser, Speech act annotation, in: K. Aijmer, C. Rühlemann (Eds.), Corpus Pragmatics. A Handbook, Cambridge University Press, Cambridge, 2015, pp. 84–114. doi:10.1017/cbo9781139057493.005.

[11] G. Leech, M. Weisser, Generic speech act annotation for task-oriented dialogues, in: D. Archer, P. Rayson, A. Wilson, T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference. Lancaster University: UCREL Technical Papers vol. 16, 2003.

[12] M. Weisser, How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond, John Benjamins, Amsterdam/Philadelphia, 2018.

[13] M. Weisser, Speech acts in corpus pragmatics: Making the case for an extended taxonomy, International Journal of Corpus Linguistics 25(4) (2020) 400–425.

[14] M. Castagneto, Il sistema di annotazione Pra.Ti.D tra gli altri sistemi di annotazione pragmatica. Le ragioni di un nuovo schema, AIΩN. Annali del Dipartimento di Studi Letterari, Linguistici e Comparati. Sezione Linguistica 1 (2012) 105–148.

[15] S. Mezza, A. Cervone, E. Stepanov, G. Tortoreto, G. Riccardi, ISO-Standard Domain-Independent Dialogue Act Tagging for Conversational Agents, in: Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, Association for Computational Linguistics, 2018, pp. 3539–3551.

[16] H. Bunt, V. Petukhova, D. Traum, J. Alexandersson Dialogue act annotation with the ISO 24617-2

standard, in: D. Dahl (Ed.), Multimodal interaction with W3C standards, Springer, Cham, 2017, pp. 109-135.
[17] E. Cresti, L. Gregori, M. Moneglia, C. Nicolás, A. Panunzi, The LABLITA Speech Resources. in E. Cresti, M. Moneglia (Eds.), Corpora e Studi Linguistici. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana, Milano, Officinaventuno, 2022, pp. 85–108.
[18] F. Cominetti, L. Gregori, E. Lombardi Vallauri, A. Panunzi, IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy (1946–2023), annotated with implicit strategies, in: D. Fišer, M. Eskevich, D. Bordon (Eds.), Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN), Torino, ELRA and ICCL, 2024, pp. 101–109.
[19] D. Archer, J. Culpeper, M. Davies, Pragmatic annotation, in: A. Lüdeling, M. Kytö (Eds.), Corpus Linguistics: An International Handbook, de Gruyter, Berlin, 2008, pp. 613–642.
[20] C. Biemann, K. Bontcheva, R. Eckart de Castilho, I. Gurevych, S. M. Yimam, Collaborative Web-Based Tools for Multi-layer Text Annotation, in: N. Ide, J. Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer, Dordrecht, 2017, pp. 229–256. doi:10.1007/978-94-024-0881-2_8.
[21] J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, I. Gurevych, The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation, in: Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING), Santa Fe, New Mexico, USA, 2018, pp. 5–9.
[22] M. A. Finlayson, T. Erjavec, Overview of Annotation Creation: Processes and Tools, in: N. Ide, J. Pustejovsky (Eds.), Handbook of Linguistic Annotation, Springer, Dordrecht, 2017, pp. 167–191. doi:10.1007/978-94-024-0881-2_5.
[23] S. Löbner, Understanding Semantics, Routledge, New York, 2013.
[24] K. Krippendorff, Content analysis: An introduction to its methodology, Sage, Thousand Oaks, 2004.
[25] K. Krippendorff, Agreement and information in the reliability of coding, Communication Methods and Measures 5 (2011) 93–112.
[26] G. C. Feng, Mistakes and how to avoid mistakes in using intercoder reliability indices, Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 11(1) (2015) 13–22.
[27] N. Duran, S. Battle, J. Smith, Inter-annotator Agreement Using the Conversation Analysis Modelling Schema, for Dialogue, Communication Methods and Measures 16(3) (2022) 182–214.
[28] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33(1) (1977) 159–174.
[29] L. Aroyo, C. Welty, Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation, AI Magazine 36 (2015) 15–24.

# Appendix A

The DIADIta annotation scheme.

| Annotation level | Tags |
|---|---|
| Forms | DISCOURSE MARKER; REPETITION; EXPRESSIVE |
| Pragmatic functions | AGREEMENT; COMMON GROUND MARKING; CONFIRMATION OF ATTENTION; DISAGREEMENT; EPISTEMICITY; EVIDENTIALITY (DIRECT, INFERENTIAL, REPORTATIVE, MEMORY); IMPOLITENESS; INTENSIFICATION; INTERRUPTION; IRONY; MIRATIVITY; MITIGATION; POLITENESS; TURN-TAKING; VAGUENESS |
| Pragmatic aims | ATTENTION-GETTING; DENIAL; DERISION; REQUEST FOR CONFIRMATION/AGREEMENT |
| Speech acts | ACCEPTANCE (OF A DIRECTIVE); ADVICE/SUGGESTION/EXHORTATION/WARNING; APOLOGY; APPROVAL/AGREEMENT; ASSERTION; CHALLENGE; COMMITMENT/PROMISE; COMPLIMENT; CONDOLENCE; CONGRATULATIONS; CORRECTION; DERISION; DISAPPROVAL/DISAGREEMENT; EXCLAMATION; FORGIVENESS; GREETING; INSULT/OFFENSE; OATH; OFFER; ORDER/COMMAND/PROHIBITION/FORBID; PERMISSION; PROPOSAL; PROTEST; REFUSAL (OF A DIRECTIVE); REPROACH/CRITICISM; REQUEST FOR INFORMATION; REQUEST FOR PERMISSION; REQUEST/PLEA; RESPONSE (TO A REQUEST FOR INFORMATION); THANKS; THREAT; WISH/HOPE |