

Communicative Intentions Annotation Scheme for Natural Language Processing Applications

María Miró Maestre¹

¹*Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain*

Abstract

Communicative intentions are one of the linguistic elements that usually determine the content of any message we want to express in our social interactions. With the purpose of contributing to the improvement of natural language processing systems, this thesis aims to create a communicative intention annotation scheme based on the taxonomy presented in the Speech Act Theory. In this way, language processing tools could consider communicative intentions as a starting point to help classify any message and its content depending first on the intention it reflects. To do so, the scheme will be created with the help of an already annotated corpus of Spanish tweets and subsequently evaluated by external annotators so that we can confirm the appropriateness and reliability of the tagged intentions before applying the scheme to an NLP system. Thus, it will be possible to check up to which point communicative intentions can improve the identification of the purpose of a message in an already created NLP system so that we can gain more linguistic information from any text automatically.

Keywords

communicative intention, annotation scheme, speech acts, natural language generation, pragmatics

1. Introduction and Motivation

Inside some of the manifold areas that comprise Natural Language Processing (NLP), the arrival of new ways of computer-mediated interaction between humans -or even humans and robots- has boosted the evolution of these technology systems. With these advances, those NLP programmes that could at first identify concrete messages inside a limited dialogue have moved on to new systems capable of adapting to different conversational contexts. These updated systems usually require more enriched structures with further linguistic knowledge to successfully fulfil research tasks such as opinion mining, sentiment analysis, or natural language generation (NLG). In order to create these NLP systems, the process generally done when including the linguistic information in the software tends to start from the lowest linguistic level of analysis and then move on to more complex levels such as semantics, if so. However, the pragmatic level of language is usually set aside given the available resources that each research project may or not have, as lower levels of linguistic analysis have an easier implementation in the system [1]. Despite this scheme adopted by most researchers when adding linguistic information to their programmes, pragmatics is starting to be considered a fundamental element


Doctoral Symposium on Natural Language Processing from the PLN.net network 2022 (RED2018-102418-T), 21-23 September 2022, A Coruña, Spain.

✉ maria.miro@ua.es (M. M. Maestre)

ORCID [0000-0001-7996-4440](https://orcid.org/0000-0001-7996-4440) (M. M. Maestre)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of analysis to successfully apply both main branches of NLP, Natural language understanding (NLU) and NLG, to real-life situations. This is mainly due to the focus of pragmatics on the study of language considering its context [2], which becomes crucial when analysing the natural component of the human language. This is due to the wide range of (para)linguistic elements that pragmatics considers, such as speakers' intentions or their previously shared knowledge, or the sociocultural context in which the message is generated, among other aspects.

Consequently, the study of pragmatics from a computational perspective has become a need that, despite the progress shown with the emergence of areas of research such as Computational Pragmatics [3], there is still a long way to go. Moreover, the diversity of research areas in which pragmatics is starting to be considered nowadays has fostered the inclusion of pragmatic aspects of language inside some of the tasks related to natural language processing, such as sentiment analysis [4], document summarisation [5] or rumour detection [6].

Pragmatics is also present in systems developed to generate natural text automatically. Indeed, inside the document planning stage, the system takes into account the type of information that needs to be included in the subsequent created text depending on factors such as the target audience or the communicative intention. However, the approaches to this type of system are usually enclosed in specific domains of application [7]. A significant proportion of NLG systems are focused on human-robot interaction so that the system can automatically understand the speaker's intentions [8], but without teaching the system how to recognise the most appropriate structure of the automatic text depending on the intention we want to determine for it.

Therefore, our motivation for the present study arises from the need for such NLP systems to include pragmatic aspects such as communicative intentions inside the tasks that focus on the structure of processing and generation systems. In this manner, we pretend to test the added value that the inclusion of this linguistic element in the structure of an NLP system could have, as it could acquire more linguistic knowledge to fulfil further linguistic tasks as a whole. To accomplish so, the main tasks to address in this thesis are the following:

- Taking as a base the taxonomy presented in the Speech Act Theory, classify a list of speech act verbs according to their main communicative intention considering their meaning
- Tag a corpus including tweets extracted from Twitter API depending on the main communicative intention they show following the previous verb classification
- Create an annotation scheme on communicative intentions taking as a model the corpus already tagged
- Apply that annotation scheme to a linguistic corpus belonging to a particular genre to test its performance
- Validate the annotation scheme by means of external annotators and evaluation metrics to confirm its reliability
- Integrate the annotation scheme and the resulting corpus in an NLP application to add further linguistic knowledge in its system, therefore increasing the tasks it can fulfil

The remainder of this article is organised as follows: Section 2 focuses on the different approaches made in NLP in order to study pragmatic elements of language, then Section 3 shows the main hypotheses and objectives planned for this research. Subsequently, we explain the methodology proposed for fulfilling each project task in Section 4, and Section 5 sets out the

different research issues that we may need to face throughout the development of the project. Finally, the bibliography used for this study is included at the end of the paper.

2. Related Work

Despite the difficulties that the inclusion of pragmatic elements inside NLP and NLG systems entailed, several researchers focused their study on this linguistic level to make progress in these domains of computational linguistics [9, 10, 11]. Therefore, there is currently a great number of studies enriching their systems with pragmatic knowledge to improve their efficiency.

A very prolific area of research is that devoted to the study of computer-mediated communication [12], which comprises all the media included in the Web 2.0, thanks to the communicative interactions it promotes with very varied tools such as Facebook or blog comments, retweets, likes on YouTube and many more. Some of the tasks studied in these types of media are analysing users' feelings, just as Tian et al. [13] did on Facebook based on the idea that emoticons reflect the intention of the message [14]. Inside the area of digital newspapers, Chen et al. [15] focused on the identification of clickbait cues using several methods of analysis that included the syntactic and pragmatic levels. As for Twitter, Saha et al. [16] and Zhang et al. [17] made use of the Speech Act Theory (SAT) founded by Austin [18] and extended by Searle [19, 20] to identify users' intentions in their tweets, modifying the intention classification with several linguistic features to apply machine learning algorithms to test their classification accuracy.

Focusing on the SAT, which we already described in detail in the previous edition of this Doctoral Symposium, its founder Austin [18] defended that language can serve as a means to perform actions depending on the uttered message, investigating verbs to identify which ones are able to denote actions on their own (called *performative verbs*) and those that only describe reality (*descriptive verbs*). Subsequent to this first pragmatic division, Austin focused his research on one of the elements that comprise the act of saying something, the illocutionary act. With this element, he created a classification divided in 5 types of illocutionary acts depending on the intention of the expressed message. However, many of the linguistic researchers that studied this theory later on took as a basis the taxonomy proposed by Searle [19], which is a more exhaustive and well delimited modification of Austin's division. According to Searle, communicative intentions can be classified in the following five categories:

- **Assertives:** by uttering them, we commit to the veracity of the message expressed. E.g.: declare, manifest, conclude, explain, etc.;
- **Directives:** the speaker uses this type to make the listener do something. E.g.: ask for, dare, invite, command, challenge, etc.;
- **Commissives:** they commit the speaker to do an action in the future. E.g.: swear, promise, commit, intend, etc.;
- **Expressives:** they express the psychological state of the speaker with respect to a topic specified in the message. E.g.: thank, forgive, excuse, congratulate, etc.;
- **Declarations:** when uttering them we get the content of the message to coincide with reality, that is, by using them, the action is performed, or in Searle's own words: '*saying makes it so*'. E.g.: declare, designate, resign, marry, etc.

Later on, Searle [20] also made a distinction between the types of intentions aforementioned, known as *direct speech acts* because the relation between the meaning and the intention of the message is clear, and other type of illocutionary acts called *indirect speech acts*. In this last type, the relation between the message and the intention requires some other inferential processes in order to successfully interpret the intention of the message —as in those messages containing irony, sarcasm or rhetorical questions, among others—. The speech act taxonomy attracted the research community in linguistics and many other fields, giving rise to different versions. More concretely, in the NLG field, this classification meant a starting point for studying the best approach to develop systems that could automatically identify text intentions [21].

Nowadays, several authors have shown interest in the SAT taxonomy, focusing on the annotation and classification of speech acts. This is the case of Martínez-Hinarejos et al. [22], who used different statistical annotation models such as the N-Gram transducer model to tag dialogue acts. Moreover, Caballero et al. [23] created a pragmatic-functional annotation scheme of the FerrovieELE corpus, with an in-depth explanation of the linguistic tags used to annotate 41 communicative functions. Focusing also on clinical pragmatics, Gallardo Paúls and Fernández Urquiza [24] applied the classification of illocutionary acts to pragmatically annotate the PerLa corpus, which contains clinical oral data to analyse pathological language.

Consequently, we base our study on these examples of pragmatic research because despite the obsolescence that SAT could show nowadays, the recognition of communicative intentions still attracts many different areas inside NLP. This is given by the general aim of processing a greater number of linguistic nuances to obtain programmes that are able to identify those linguistic particularities and therefore process a wider variety of text genres considering the pragmatic nature of language.

3. Main hypothesis and objectives

The present thesis is based on the hypothesis that it is possible to automatically annotate the communicative intention of a given message by means of a representative and unambiguous annotation scheme. Consequently, the purpose of this project is the creation of a communicative intention annotation scheme that could serve as a model for the pragmatic annotation of several NLP applications to broaden the linguistic scope of this area of research. By establishing an annotation manual adaptable to different research purposes, NLP and NLG systems could be further trained with pragmatic information to understand and classify different texts depending on the particular intention they reflect. In this way, the improvement of those computational systems with more heterogeneous information will foster the creation of processing systems with more of the linguistic elements that make a text look natural, and therefore achieve one of the multiple purposes of these research areas. To tackle this pragmatic subject within our PhD research, we aim to answer the following research questions:

- Up to which point is it possible to identify the intention of a given message in Spanish?
- What NLP tools do we need to process and detect those intentions automatically?
- How to evaluate the annotation scheme to validate its effectiveness?
- In which NLP application should we implement our scheme to check its performance?

4. Methodology and proposed experiment

The proposed research is based on the application of the SAT in an annotation scheme that could serve as a model for future NLP systems in order to automatically recognise and tag the communicative intention of a particular message in a given corpus or application. Consequently, for the purpose of our thesis, we will focus on Searle's classification of *direct speech acts* as explained in Section 2 and some other linguistic features that also reflect the intention of the message in a straightforward way. To create the corresponding annotation scheme, several linguistic resources and computing tools were used to collect the sufficient linguistic information that would serve as the base of the guidelines:

- Anne Wierzbicka's *English Speech Act Verbs: A Semantic Dictionary* [25]

After consolidating the theoretical foundations of the SAT in the previous Doctoral Symposium, the next step of the thesis was selecting an appropriate lexicon that contains a considerable representation of the verbs comprised in Spanish to then include them in the annotation scheme according to their essential communicative intention. In order to have a clear idea of which verbs were going to be included in our classification of speech act verbs, we based our selection on the verbs semantically analysed in [25]. This book describes in detail around 200 of the most frequently speech act verbs used in the English language focusing on their particular semantic meaning. In this way, by studying the semantic particularities of each English verb, we were capable of looking for the equivalent verbs in Spanish that kept each semantic nuance so that the speech act verb classification would not differ from one language to another.

- ADESSE: Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español [26]

Along with the speech act classification described in [25], we also used for the creation of our verb classification the online database ADESSE [26], created by the University of Vigo. This online linguistic tool shows a Spanish verb and two-verb constructions database that includes a syntactic and semantic analysis of those verbs inside the corpus Arthus. The corresponding corpus contains 34 texts belonging to the narrative genre, including newspapers and theatre plays, among others. Furthermore, ADESSE incorporates an scheme of the semantic classes in which each verb analysed was included, which was of great help when matching the semantic senses of the verbs analysed in [25] with those of the ADESSE database in Spanish, so we could support our classification with official linguistic resources.

- Shared Task on Hope Speech Detection for Equality, Diversity and Inclusion [27]

At the same time, a corpus of an adequate length was chosen to identify messages with a particular communicative intention depending on the verb classification previously completed. For this task, we first adopted the corpus published by the shared task competition on Hope Speech Detection for Equality, Diversity and Inclusion [27]. This shared task includes a Spanish dataset of LGBTQ-related tweets collected using the Twitter API from June 27, 2021 to July 26, 2021. Even though the purpose of the competition was to check the performance of the

different systems participating in the task when tagging the tweets that included or not hope speech messages, we found this dataset quite suitable for our own research. This is because many of the tweets gathered in the train dataset used several of those speech act verbs that we had previously included in our classification.

- Twitter API

Nevertheless, as we started to tag the tweets depending on their intention, we realised that we could not classify as many tweets as we considered appropriate to verify that our verb classification was correct. Also, with the reduced number of tweets classified with this corpus, we could not identify as many linguistic features that could also mark the communicative intention of the message without ambiguities. For this reason, we extracted more tweets through Twitter API using the same hashtags than in [27] to complete our first corpus of tweets classified depending on their main communicative intention to gather more linguistic features.

4.1. Communicative intention annotation scheme

With the linguistic tools and the communicative intention classification already established, the next part of our research, which is currently under development, focuses on creating the annotation scheme. In this step of the thesis, we will gather the speech act verbs classification apart from some other linguistic features found throughout the tagging of the tweets, which also helped to a large extent to identify the central intention of our compiled tweets. Some of these features linked to the Spanish language are verbal periphrases as "*deber + infinitive verb*" (which would denote an obligation) and fixed Spanish expressions that substitute the meaning of a particular verb in English, as in "*agree*" in comparison to "*estar de acuerdo*", among some other features. The annotation scheme will also count with a section devoted to the parameters and rules to consider for the intention annotation task of the linguistic features included in our scheme. To ensure a clear and simple annotation performance, the scheme will show the most representative usage examples of each intention so that, wherever possible, we avoid the ambiguities expected in some annotated cases.

Apart from the examples added in each intention to illustrate its annotation, as shown in Table 1, the proposed annotation scheme will also contain a final glossary with every annotated verb and its corresponding intention as a result of the previous annotation task. In this way, the annotation scheme will also become a linguistic resource in future studies that want to focus on the intentions applied to many areas within the field of NLP. Moreover, the annotation scheme will have a technical section devoted to the tags chosen for the annotation task so that this stage is performed in the most visual and mechanical way possible to create an effective annotation system that can be implemented in an NLP programme. This technical section will also be crucial for assessing the quality of our annotation scheme, as external annotators will be needed in order to verify its accuracy. To do so, a special section devoted to evaluation metrics and the inter-annotator agreement will also be included in the thesis to corroborate the reliability of the communicative intention annotation scheme.

- Experimentation

Table 1
Tweet annotation example

Original tweet	Annotated tweet
En nombre del colectivo #LGTBI de #Chivilcoy agradecemos siempre la predisposición a la escucha y a la acción para hacer un municipio cada día más justo	<expres> En nombre del colectivo #LGTBI de #Chivilcoy <vah_expres> agradecemos </vah_expres> siempre la predisposición a la escucha y a la acción para hacer un municipio cada día más justo </expres>
FELIZ DIA DEL ORGULLO LGBTQ+ y a los HETEROS ALIADOS les recordamos , ustedes también forman parte de nosotros, gracias por apoyar. #Orgullo2021 #pride #Orgullo #OrgulloSiempre #OrgulloLGTBI #LGBT	<expres> <fra> FELIZ DIA </fra> DEL ORGULLO LGBTQ+ </expres> <repre> y a los HETEROS ALIADOS les <vah_repre> recordamos </vah_repre>, ustedes también forman parte de nosotros, </repre> <expres> <fra> gracias </fra> por apoyar. #Orgullo2021 #pride #Orgullo #OrgulloSiempre #OrgulloLGTBI #LGBT </expres>

Finally, we will proceed to the experimentation of the annotation scheme in an NLP system once it is validated. The real value of these guidelines is the variety of applications in which they can be tested to broaden the number of actions an NLP system can successfully complete. In this manner, our pragmatic scheme could help an NLG system to generate automatically created text with a particular intention stated in it, which would help to a great extent to structure the rest of the information we want to represent in the generated text. Apart from this, the annotation scheme could be included in an already developed NLP application devoted to the task of opinion mining. In this case, the scheme would be implemented as another section of the application so that the system can recognise not only the opinion of a given message, but also its main intention, as well as many other (para)linguistic aspects that would increase the number of tasks that this particular system is capable of fulfilling. Furthermore, another lines of future research could be focused on the application of our annotation guidelines into another language or analysing which other pragmatic features could be also annotated in the corpus resulting from our communicative intention annotation task.

5. Research issues to discuss

Given the suggestions and comments received in the previous edition of the Doctoral Symposium, we solved some of the research issues stated at the beginning of the thesis. However, as an inherent part of this project, there are still several research questions to be discussed all throughout the development of our annotation scheme:

- Should we avoid annotating Declarative verbs as they depend on more linguistic and contextual information in order to be classified as such? (i.e., Who is expressing that verb (his job), and to whom it is said, etc.)
- How much linguistic information is it interesting to include in the annotation scheme apart from the speech act verbs to not worsen its performance?

- How many tweets are enough to check all the possible speech act verbs and other linguistic features that may be added to the subsequent annotation scheme?
- How are we going to tackle the particularities of a computer-mediated type of communication so singular as tweets? Spelling mistakes, punctuation mistakes, emojis, etc.

Acknowledgments

This research work is part of the R&D project "PID2021-123956OB-I00", funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe". Moreover, it has been partially funded by the Generalitat Valenciana through the project NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)".

References

- [1] C. Cherpas, Natural language processing, pragmatics, and verbal behavior, *The Analysis of Verbal Behavior* 10 (1992) 135–147. doi:10.1007/bf03392880.
- [2] R. Resende de Mendonça, D. Felix de Brito, F. de Franco Rosa, J. C. dos Reis, R. Bonacin, A framework for detecting intentions of criminal acts in social media: A case study on twitter, *Information* 11 (2020) 1–40. doi:10.3390/info11030154.
- [3] D. Sayers, R. Sousa-Silva, S. Höhn, L. Ahmed, K. Allkivi-Metsoja, D. Anastasiou, Š. Beňuš, L. Bowker, E. Bytyçi, A. Catala, et al., The Dawn of the Human-Machine Era: A Forecast of New and Emerging Language Technologies, Technical Report, EU COST Action, 2021. URL: <https://hal.archives-ouvertes.fr/hal-03230287>.
- [4] T. Mahler, W. Cheung, M. Elsner, D. King, M.-C. de Marneffe, C. Shain, S. Stevens-Guille, M. White, Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems, in: *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, 2017, pp. 33–39.
- [5] B. A. Mukhedkar, D. Sakhare, R. Kumar, Pragmatic analysis based document summarization, *International Journal of Computer Science and Information Security* 14 (2016) 145–149.
- [6] A. Kumar, S. R. Sangwan, A. Nayyar, Rumour veracity detection on twitter using particle swarm optimized shallow classifiers, *Multimedia Tools and Applications* 78 (2019) 24083–24101. doi:10.1007/s11042-019-7398-6.
- [7] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, *Journal of Artificial Intelligence Research* 61 (2018) 65–170. doi:10.1613/jair.5477.
- [8] K. Garoufi, Planning-based models of natural language generation, *Language and Linguistics Compass* 8 (2014) 1–10. doi:10.1111/lnc3.12053.
- [9] W. C. Mann, *Toward a Speech Act Theory for Natural Language Processing*, Technical Report, University of Southern California Marina del Rey Information Science Inst, 1980.
- [10] S. C. Herring, D. Stein, T. Virtanen, Introduction to the pragmatics of computer-mediated communication, in: *Pragmatics of Computer-Mediated Communication*, De Gruyter Mouton, 2013, pp. 3–32. doi:10.1515/9783110214468.

- [11] C. Bonial, L. Donatelli, M. Abrams, S. Lukin, S. Tratz, M. Marge, R. Artstein, D. Traum, C. Voss, Dialogue-amr: abstract meaning representation for dialogue, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 684–695.
- [12] A. Georgakopoulou, Computer-mediated communication, in: J. Verschueren, J.-O. Östman, J. Blommaert, C. Bulcaen (Eds.), *Pragmatics in Practice*, volume 9, John Benjamins Publishing Co, 2011, pp. 93–110.
- [13] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, C. Sun, Facebook sentiment: Reactions and emojis, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, ACL, 2017, pp. 11–16. doi:10.18653/v1/W17-1102.
- [14] E. Dresner, S. C. Herring, Functions of the nonverbal in CMC: Emoticons and illocutionary force, *Communication Theory* 20 (2010) 249–268. doi:10.1111/j.1468-2885.2010.01362.x.
- [15] Y. Chen, N. J. Conroy, V. L. Rubin, Misleading online content: recognizing clickbait as "false news", in: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection, ACM, 2015, pp. 15–19. doi:10.1145/2823465.2823467.
- [16] T. Saha, S. Saha, P. Bhattacharyya, Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8. doi:10.1109/IJCNN.2019.8851805.
- [17] R. Zhang, D. Gao, W. Li, What are tweeters doing: Recognizing speech acts in twitter, in: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011, pp. 86–91. URL: <https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3803>.
- [18] J. L. Austin, *How to Do Things with Words*, Oxford at the Clarendon Press, 1962.
- [19] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*, volume 626, Cambridge University Press, 1969.
- [20] J. R. Searle, *Expression and meaning: Studies in the theory of speech acts*, Cambridge University Press, 1985.
- [21] G. Briggs, M. Scheutz, A hybrid architectural approach to understanding and appropriately generating indirect speech acts, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 27, 2013, pp. 1213–1219. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/8471>.
- [22] C. D. Martínez-Hinarejos, J. M. Benedí, V. Tamarit, Unsegmented dialogue act annotation and decoding with n-gram transducers, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (2014) 198–211. doi:10.1109/TASLP.2014.2377595.
- [23] M. Caballero, L. Díaz, M. Taulé, *Guía de anotación del corpus FerroviELE*, 2014.
- [24] B. Gallardo Paúls, M. Fernández Urquiza, Etiquetado pragmático de datos clínicos, *e-AESLA* (2015) 1–12.
- [25] A. Wierzbicka, *English Speech Act Verbs: A Semantic Dictionary*, Academic Press, 1987.
- [26] J. M. García-Miguel, F. González Domínguez, G. Vaamonde, I. Anaya, A. Huzum, V. Dacosta, A. Rifón, ADESSE: Base de datos de Verbos, *Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*, 2010. URL: <http://adesse.uvigo.es/index.php>.
- [27] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. C. Navaneethkrishnan, J. P. McCrae, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, R. Valencia-García, Shared task on hope speech detection for equality, diversity, and inclusion - ACL, 2022. URL: https://competitions.codalab.org/competitions/36393#learn_the_details-organizers.