Scientific Research Publishing

Open Journal of Modern Linguistics, 2019, 9, 206-214
http://www.scirp.org/journal/ojml
ISSN Online: 2164-2834
ISSN Print: 2164-2818

A Simple Framework for the Annotation of Small Corpora

David González Gándara

Universidade de Santiago de Compostela, Santiago de Compostela, Spain
Email: david.gonzalez.gandara@rai.usc.es

How to cite this paper: Gándara, D. G. (2019). A Simple Framework for the Annotation of Small Corpora. *Open Journal of Modern Linguistics, 9,* 206-214.
https://doi.org/10.4236/ojml.2019.93019

Received: May 4, 2019
Accepted: June 16, 2019
Published: June 19, 2019

Abstract

This paper describes a framework for the annotation of discourse which consists of the combination of software tools and a tagset. Its main purpose is to create small corpora for action-research projects conducted by Second of Foreign Language teachers, including Content and Language Integrated Learning, in their classrooms. The framework intends to be a solution for this particular situation, which requires a high level of simplicity. Although in the literature of Corpus Linguistics there are some good frameworks and tagsets for the annotation of corpora, they are usually too complex for teachers who are not experts either in Corpus Linguistics or Discourse Analysis.

Keywords

Discourse Analysis, Corpus Linguistics, Manual Tagging, Classroom Discourse, Computational Linguistics, Pragmatics Tagging

1. Introduction

I started teaching English twenty years ago. With the purpose of becoming a better teacher, I observed other teachers' work to find new ideas, read books and went to conferences. However, I quickly found that was not enough. Being able to do a systematic observation and assessment of my own teaching was necessary, and action-research seemed to be the way to do it. However, administrators in Spain do not plan any training for teachers to learn how to do research. There is the possibility to enroll in research projects at universities, but in this context there is a tendency for the use of big data and very specialized and complex knowledge.

What teachers really need to improve their own practice is small-scale projects and the simplification of concepts and processes. Sometimes because they are not really interested, and sometimes because they do not have the time. They

teach full time.

With regards to the specific characteristics of Second and Foreign Language classrooms, the framework here proposed aims at the observation of language use in communicative tasks, which involves pragmatics. A consequence of this is that annotations have to be completed manually. The field of automated annotation—which is particularly useful for large corpora—has developed to a state of great accuracy and reliability with regards to parts of speech (POS). However, for more complex linguistic concepts, such as semantics or pragmatics, manual annotation is still needed.

Nevertheless, manual annotation of small corpora can be the basis for training algorithms that, in the future, will be able to perform more complex tasks with the same accuracy (Hovy & Lavid, 2010).

This paper is structured following the stages of the framework. These stages are: recording the sessions, transcribing the interactions in communicative tasks, annotating the transcriptions, and analyzing the data.

2. Stage 1: Recording the Sessions

Recording the lessons is a necessary step in order to aim research at discourse analysis, which is one of the areas I became interested. Today it became very easy to obtain high quality audio without having to buy expensive equipment or high skill about audio recording. Microphones in computers, mobile phones or tablets offer quality enough to allow a smooth transcription process. Technical specifications of audio recording fall out of the scope of this paper.

3. Stage 2: Transcribing

This stage can begin as soon as the action-researcher has some recordings available from stage 1, so that both stage 1 and 2 may be performed simultaneously. It is in this part of the process that this framework comes into action. In order to complete this stage I wrote an *Emacs* package (Lewis, LaLiberte, & Stallman, 1990) called *Transcribe*, written in *Emacs Lisp* (a programming language) by myself, which provides with key combinations to do some operations over the audio files on the same screen that the user writes the transcriptions (**Figure 1**): start, stop, go forward, go backwards. This is achieved by the common *Emacs* package EMMS, which can be found in the project's web page. The package *Transcribe* can be easily installed from the *Emacs* package repositories. Once the package is installed, it is easy to get information about the keys that perform the operations, and they can be customized. It is a free software package, so any part of it can be modified to the needs of the user. It provides the *transcribe-mode* in *Emacs*, which will add the menu *Transcribe* in the menu bar, with the different actions that can be performed.

The user may just write the transcriptions from the audio and save the file, but I recommend to add at least some of the annotation tags at the same time in the manner that is going to be described in the next sections. This method will save a lot of time if the transcriber and the annotator is the same person.

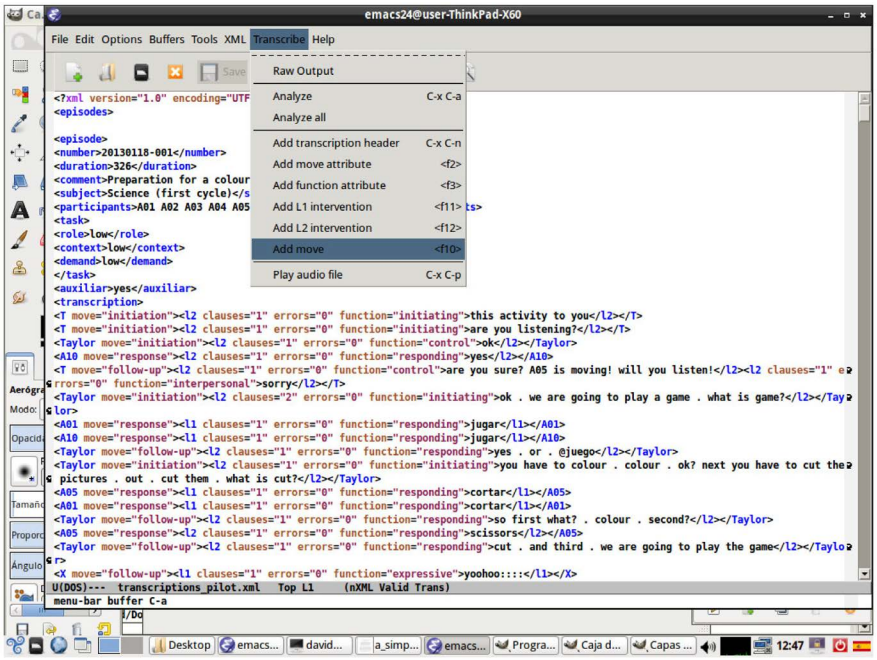

Figure 1. *Transcribe-mode* in action.

4. Stage 3: Annotation of Discourse and Pragmatics Features

When the transcription stage is finished, the action-researcher can annotate the text. I recommend to complete all the transcriptions before beginning stage 3 because it simplifies the planning of the annotating process.

For the annotation of discourse and pragmatic features, it is necessary to define a scheme. What I propose is a set of XML (Extended Markup Language) tags. As a basis to create a very simple scheme that can be applied by people who are not expert in linguistics I revised some schemes that I found in previous literature.

As a first example, the functions described in DASML (Allen & Core, 1997; Core & Allen, 1997; Jurafsky, Shriberg, & Biasca, 1997) are too many in my view (statements: assert, reassert, other-statement; influencing-addresee-future-action: open-action, action-directive; info-request; commiting-speaker-future-action: offer, commit; conventional: opening, closing; explicit performative; exclamation; other-forward-function). Apart of the eight categories, it includes some subcategories. Besides the number of categories, it appears to be complicated to decide which one to apply, the names require a deep knowledge of linguistics. This is a problem for the person who is transcribing. If the categories are too narrow and hard to understand classifying is too large an effort for teachers who are using their free time for research.

As for SPAAC (Leech & Weisser, 2003), I found in this scheme really interesting features. In fact, I based my scheme mostly on ideas from this model. In particular, the taxonomy for speech acts. However, the original scheme becomes too complex at the lowest level of concretion, even the authors report problems deciding which tags to apply for each situation. The model describes more than

40 types of speech acts, whereas in my model there are only 5 categories because the lowest level is not used.

In the case of PRAGMATEXT (González-Ledesma, 2007), the problem that I found is that the intention of the scheme was way more ambitious than mine. It is intended to take into account features of discourse such as emotional discourse, modalization, evidentiality, phraseology or metaphor. As a result, it brings many annotation elements which are completely unnecessary for the kind of projects my model is aimed.

With regards to EXMARaLDA (Schmidt & Wörner, 2009), it is aimed to automatic software processing of the files. From the beginning, I wanted a system which does not depend on specific software and that could even be edited manually. EXMARaLDA does not provide this.

I also analysed the scheme designed by Bunt et al. (2010), which tries to go further and define the basics for an ISO compliant markup system.

Especially interesting is the work of Chiarcos et al. (2008), which establishes a framework to integrate different types of annotation schemes, and set some standards that the schemes have to comply with in order to allow its integration with others. However, the standard they propose is also unnecessarily complicated for small scale studies. This can be observed in this example of a tag: "<mark id="tok_21" xlink:href="#xpointer(string-range(//body,",101,4))"/>" (Chiarcos et al., 2008), which is not precisely user friendly. Nonetheless, the possibility of translating from one standard to another is really interesting. The way I designed the scheme makes it possible for easy translation, by means of a script into the PAULA standard defined in the mentioned work (Chiarcos et al., 2008).

All the schemes analyzed were designed either with too many different tags or their technical specifications were too complex to be performed by untrained teachers. If any of these schemes had reached the status of standard, it could compensate their complexity. Unfortunately, even though some of these schemes are widely used, none of them has become a clear standard yet.

In contrast with the complexity of the models found in previous literature, the new scheme that I propose is simple enough for anyone who reads the instructions. The first step is to create the file that will store the transcriptions. Some information is needed in the header of the file, working as metadata, so that the automatic processing of the transcriptions is easier. After the required line for a XML file, the tag <episodes> will encapsulate some <episode> sub-tags (**Listing 1**). This is the highest layer of the scheme proposed here. The following tags give more information about each <episode> (*The Transcribe* package described above allows automation of this):

<number> The function of this tag is to provide a label to identify the episode.

<duration> In this tag the duration of the episode is given in seconds.

<comment> Any comments that the transcriber or the annotator want to add.

<subject> The subject of the lesson in which the episode took place.

```
<? xml version = " 1.0 " encoding = " UTF -8 " ?>
<episodes>
<episode>
<number> 20130118 -001 </number>
<duration> 326 </duration>
<comment> Preparation for a colouring activity </comment>
<subject> Science (first cycle) </subject>
<participants> A01 A02 A03 A04 A05 A06 A07 A08 A09 A10 X T </participants>
<task>
<role> low </role>
<context> low </context>
<demand> low </demand>
</task>
<auxiliar> yes </auxiliar>
<transcription>
...
</transcription>
</episode>
</episodes>
```

Listing 1. Structure of the transcription file.

<participants> A list of the identification numbers that are going to be used in the transcription.

<task> Each episode constitutes a task that took part in the lesson. Inside this tag, the sub-tags: <role>, <context> and <demand> can be defined. These sub-tags correspond to the task taxonomy proposed in González Gándara (2017) and González Gándara (2019), where some examples of the use of this scheme can be found.

The database file format is structured in layers that can be added on top of each other indefinitely. The lowest level layer is composed by the speech acts. The segmentation will be done in terms of AS-Units (Analysis of Speech Units) (Foster, Tonkyn, & Wigglesworth, 2000). As-units are defined as: "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either". These As-Units are tagged "<l1>" or "<l2>", depending on the language that is used (**Listing 2**). The annotator may add attributes to this tags that can be analysed by the statistical software (My recommendation is *R*). In my research, I use the tags:

<clauses> The number of clauses in the speech act.

<errors> The number of error in the speech act.

<function> The pragmatic function of the speech act.

The number of clauses is an objective measure, the annotator just have to count.

```
<l2 clauses="1" errors="0" function="responding"> school? </l2>
<l2 clauses="1" errors="0" function="responding"> of course school . </l2><l2 clauses="1" errors="0" function="expressive"> how clever .. </l2>
```

Listing 2. Example of speech acts.

In the case of the errors, this is more subjective, and it will depend on the purpose of the researcher. As for the linguistic function, the superordinate clauses defined by Leech & Weisser (2003) offer the simplicity that I am looking for this scheme:

Initiating direct, request information, suggest, inform, etc.

Responding accept, acknowledge, answer, confirm, negate, etc.

Dialogue Control complete, correct, correct-self, echo, identify, pardon, etc.

Expressive exclaim, express opinion, express possibility, express regret, express wish, etc.

Interpersonal Management thank, greet, bye, etc.

I found these much easier to understand to non-experts than other classical taxonomies of speech acts (Austin, 1962; Searle, 1976). Besides, this model has been incorporated in widely used corpora, like the British National Corpus or the Lancaster-Oslo-Bergen Corpus.

On top of the speech act level (As-units), the next layer consists of the "moves". To annotate the moves I propose the classical "IRF" model (Sinclair & Coulthard, 1975) which defines the moves: initiation, response and follow up as the typical structure of classroom discourse. For this purpose, the tags of the former layer will be introduced inside a mother tag that defines the speaker and the move (**Listing 3**).

I decided not to include a specific layer to encapsulate the IRF exchanges because it creates a high level of complexity, in terms of deciding how to set the limits of each exchange. Sometimes they overlap, sometimes they are incomplete, etc. Apart from this, the fact of annotating the limits of each exchange does not offer interest enough for the analysis of discourse to be worth the effort.

5. Stage 4: Analysis of Data

When stages 1 to 3 are completed, the action-researcher may start the analysis of data. One important feature that the package *Transcribe* provides is the possibility to get automatically a text file (**Listing 4**) that can be analyzed by statistical software.

The output provided is design to work in *R* scripts (Team, 2008). For backward compatibility, the actual package provides with measures I no longer use in my own research, but they might be of interest for other researchers. They could also be removed if they are not used in the *R* scripts. The measurements that *Transcribe* is able to perform out of the package are:

QUAN1 The amount of as-units per second in the L1.

QUAN2 The amount of as-units per second in the L2.

```
<T move="initiation"><l2 clauses="2" errors="0" function="control"> please can you listen instead
  of speaking .. </l2><l2 clauses="1" errors="0" function="initiating"> whats your favourite game .
  </l2><l2 clauses="1" errors="0" function="initiating"> for example . my favourite game is chess .
  </l2><l2 clauses="1" errors="0" function="initiating"> chess is ajedrez . </l2><l2 clauses="1"
  erors="0" function="initiating"> whats your favourite game ? </l2></T>
<A13 move="response"><l1 clauses="1" errors="0" function="responding"> de que ? </l1></A13>
<T move="follow-up"><l2 clauses="1" errors="0" function= "initiating"> game .
your favourite game </ l2 ></ T >
```

Listing 3. Typical IRF exchange.

```
person,episode,duration,C-UNITS(L2),C-UNITS(L1),role,context,demand,        QUAN-L2,QUAN-
  L1,QUAL-L2,initiating,responding,control,expressive,interpersonal,shifts,
  aux,level,subjects,yearofCLIL,month
A10,20130118-001,326,0,0,low,low,low,0.015337423312883436,0.009202453987730062,
QUAL-L2,0.0,1.0,0.0,0.0,0.0,1.0,aux,level,subject,yearofclil,month
```

Listing 4. Text output prepare for statistical analysis in *R*.

Initiating The number of "initiating" As-units divided by the total of As-units in the episode.

Responding the number of "responding" As-units divided by the total of As-units in the episode.

Control the number of "control" As-units divided by the total of As-units in the episode.

Expressive the number of "expressive" As-units divided by the total of As-units in the episode.

Interpersonal the number of "interpersonal" As-units divided by the total of As-units in the episode.

Shifts the number of shifts from L2 to L1 per As-unit.

However, other measurements can be added to the package by modifying the source code, which is provided.

6. Conclusion

This framework was used to build the corpus of my lessons in the context of my own research (González Gándara, 2017, 2019). However, in this paper, I tried to demonstrate that it can be a contribution to foster new small action-research projects which will enrich the literature about communicative tasks in Second or Foreign Language Teaching, including Content and Language Integrated Learning (CLIL).

The combined use of the EMMS and the *Transcribe* packages in *Emacs* with the statistical software *R* and the annotation scheme proposed in this paper provides a very accessible way for second of foreign language teachers who want to improve their training by performing action-research projects about their own lessons, attain sound and well-founded analysis of discourse in their classrooms

and come to rich conclusions.

Although the package *Transcribe* was designed for a specific research study, it is a piece of free software, so it can be easily adapted to other conditions as well as being exported to comply with standards like *PAULA* (Chiarcos et al., 2008).

Besides its function as a tool for action-research, small annotated corpora created with the method presented in this paper may be used for the training of algorithms for automatic annotation of pragmatics.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog Act Markup in Several Layers.* Rochester: University of Rochester.

Austin, J. L. (1962). *How to Do Things with Words.* London: Oxford University Press.

Bunt, H., Alexandersson, J., Carletta, J., Choe, J.-W., Fang, A. C., Hasida, K., Romary, L. et al. (2010). Towards an ISO Standard for Dialogue Act Annotation. In *Seventh Conference on International Language Resources and Evaluation* (pp. 2548-2555). Valetta, Malta.

Chiarcos, C., Dipper, S., Goetze, M., Lesser, U., Luedelin, A., Ritz, J., & Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitment Automatique des Langues, 49,* 217-246.

Core, M. G., & Allen, J. (1997). Coding Dialogs with the Damsl Annotation Scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines* (Vol. 56, pp. 28-35). Cambridge, MA.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring Spoken Language: A Unit for All Reasons. *Applied Linguistics, 21,* 354-375. https://doi.org/10.1093/applin/21.3.354

González Gándara, D. (2017). The Role of the Students in the CLIL Classroom a New Perspective to Identify Types of Tasks. *International Journal of Applied Linguistics and English Literature, 6,* 5-10. https://doi.org/10.7575/aiac.ijalel.v.6n.4p.5

González Gándara, D. (2019). *Discourse Analysis in the CLIL Classroom. The Effects of Task Dimensions on L2 Oral Performance.* PhD Thesis, Santiago de Compostela: Universidade de Santiago de Compostela.

González-Ledesma, A. (2007). *PRAGMATEXT: Annotating the C-ORAL-ROM Corpus with Pragmatic Knowledge*. Birmingham: University of Birmingham.

Hovy, E., & Lavid, J. (2010). Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation, 22,* 13-36.

Jurafsky, D., Shriberg, L., & Biasca, D. (1997). *Switchboard SWBD-DAMSL Shallow Discourse-Function Annotation: Coders Manual* (pp. 97-102). Institute of Cognitive Science Technical Report. Boulder, CO, USA: University of Colorado.

Leech, G., & Weisser, M. (2003). Generic Speech Act Annotation for Task-Oriented Dialogue. In D. Archer, P. Rayson, A. Wilson, & A. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 441-446). Lancaster: UCREL Technical Pa-

pers.

Lewis, B., LaLiberte, D., & Stallman, R. (1990). *GNU EMACS LISP Reference Manual*. Cambridge, MA: Free Software Foundation.

Schmidt, T., & Wörner, K. (2009). EXMARalDA-Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. *Pragmatics. Quarterly Publication of the International Pragmatics Association, 19,* 565-582. https://doi.org/10.1075/prag.19.4.06sch

Searle, J. R. (1976). A Classification of Illocutionary Acts. *Language in Society, 5,* 1-23. https://doi.org/10.1017/S0047404500006837

Sinclair, J., & Coulthard, M. (1975). *Towards an Analysis of Discourse. The English Used by Teachers and Pupils*. Oxford: Oxford University Press.

Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. http://www.R-project.org