Corpus tools for parallel corpora of theatre plays: an introduction to TAligner and ACM-theatre

Olaia Andaluz-Pinedo*
University of the Basque Country UPV/EHU
olaia.andaluz@ehu.eus
0000-0002-9975-4464
*Corresponding author

Hugo Sanjurjo-González
University of Deusto
hugo.sanjurjo@deusto.es
0000-0001-5874-9733

Abstract Software tools are of vital importance in corpus-based research, but they can also lead to restrictions on the type of supported corpora and the range of analyses that can be performed. For example, corpus analysis tools, as general purpose software, do not include specific features to process corpora of theatre plays. This situation is even worse for parallel corpora of theatrical texts, in that there is currently a lack of software that allows for both the alignment and analysis of parallel corpora here. In this contribution, we will first outline the peculiarities of theatre texts and suggest three software features to address them: annotation of the structural units of plays, alignment at the utterance level, and concordances and statistics using the annotated units. Second, we will present the specific functionalities of TAligner and ACM to build and analyse parallel corpora of play texts, showing how new avenues of research are opening up with the development of these tools.

Keywords corpus building, corpus analysis, software, parallel corpora, theatre translations

Declarations

Funding: Part of this study was funded by the Spanish Agency for Research, Development and Innovation (Ministry of Economy and Competitiveness) [FFI2016-75672-R]. At the time of writing, the co-author Olaia Andaluz-Pinedo is a doctoral student funded by the University of the Basque Country UPV/EHU, Spain.

Conflicts of interest: The authors declare that they have no conflict of interest.

Availability of data and material: Not applicable

Code availability: Software application

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

1. Introduction

There is broad consensus regarding the usefulness of parallel corpora[1] for contrastive linguistics and descriptive translation studies, as well as for translation practice, the training of translators, lexicography, and foreign language teaching (Doval and Sánchez-Nieto 2019: 3). Since the development of the first well-known parallel corpus, the English-Norwegian Parallel Corpus (Johansson and Hofland 1994; Oksefjell 1999), several corpora of this type have been created, including the European Parliament Corpus (Koehn 2005), UN Parallel Corpus (Rafalovitch and Dale 2009), P-ACTRES 2.0 (Sanjurjo-González and Izquierdo 2019), ALEUSKA (Sanz-Villar 2019), COVALT (Marco 2019; Molés-Cases and Oster 2019) and MULTINOT (Lavid 2019), the growing availability of such resources reflecting an increasing interest in parallel corpora studies.

Parallel corpus building and analysis are typically carried out using a variety of different software tools. As we know, software for the analysis of corpora opens the door to the use of a range of helpful analytical techniques, such as concordances, qualitative and quantitative statistics, and visualisations. While such tools provide users with invaluable help, it is also true that they pose limits regarding what in fact can be done within certain areas of study (McEnery and Hardie 2012: 36; Anthony 2013: 146-147). Hence, Anthony (2013: 141-151) has made a case for the importance of differentiating between corpora and tools, and being aware of the influence of the latter on the scope of possible research of the former.

An interest in tools as a key element in corpus-based studies is at the core of the present contribution. In particular, we will address the absence of software available for parallel corpora that can accommodate structural features of texts from a particular genre: drama. This issue seems to have been largely overlooked in corpus analysis software (Sanjurjo-González 2018: 47-48). While available software caters for corpora of many text types written in prose, there is no reason why theatre corpora should miss out on the potential advantages that such tools undoubtedly offer. In this sense, the adjustment of certain technological aspects relating to the structure of texts would go a long way towards facilitating research on parallel theatre corpora.

Parallel corpora of plays present two main challenges from the perspective of corpus creation and analysis using tools. One of these issues is related to the requirements of parallel corpora, and the other to the characteristics of theatre texts themselves. On the one hand, there are not many tools that support parallel corpora analysis, and even fewer that integrate parallel corpus building, which involves alignment (Sanjurjo-González 2018: 47-48). AntPConc (Anthony 2014), Sketch Engine (Kilgarriff et al. 2014) and CQPweb (Hardie 2012) allow analysis of aligned corpora in some way, but alignment has to be performed externally (Sanjurjo-González 2018: 47-48). ACM

---

[1] Following Xiao and Yue (2004: 240), parallel corpora are understood as a set of source texts aligned with their translations. From the perspective of corpus tools, parallel corpora are more complex than monolingual or comparable ones, since issues such as alignment need to be considered (Sanjurjo-González 2018: 25). Although the focus of this contribution is on parallel corpora, the adjustments of tools for structural annotation and specific analytical functions may also be useful for monolingual corpora.

(Sanjurjo-González 2017a) fills this gap by providing a framework that offers both automatic alignment and analysis features. Similarly, TAligner[2] integrates manual alignment with a querying interface (Sanz-Villar and Andaluz-Pinedo 2021). On the other hand, available tools do not support specific features required for parallel corpora of plays (see section 2). Recently, the existing versions of TAligner and ACM-theatre have been adapted to support the peculiarities of theatre corpora.

In the remainder of this study we will explore the potential of software to build and analyse parallel corpora of theatre plays. First, the structural peculiarities of such texts and the implications in terms of software design are discussed. Second, we present two applications adapted to create and analyse parallel corpora of plays: TAligner and ACM-theatre. These tools offer different features, which will be compared. Finally, we will present conclusions.

2. Building and analysing a parallel theatre corpus

The organisation of a dramatic text in its relatively complex set of structural units is specific to the "field of drama" (Esslin 1990; Merino-Álvarez 1994: 44-46), hence differentiating the genre from both prose and poetry.[3] The processing of these units in corpus tools is useful in order to build and analyse parallel corpora of theatre plays. However, few tools are able to do so. In this section, we will suggest some helpful software features, based on the structural specificity of plays and a review of previous work on theatre corpora.

Merino-Álvarez (1994: 44-46), in her research on theatre translations, notes a useful systematisation of the units in which dramatic texts are normally structured: acts/scenes, utterances, speakers, stage directions and dialogue. Global units such as acts or scenes usually divide theatre texts in a similar way to how chapters organise narrative texts (Merino-Álvarez 1994: 45). The utterance, inherent to theatrical works, is seen by Merino-Álvarez (1992: 285, 1994: 44-46) as the minimal structural unit of the genre. Speakers, stage directions and dialogue are further structural subdivisions of utterances: "each utterance is clearly indicated in the page by the name of the character which tells us when the turn for the said character to speak (and move) has come" (Merino-Álvarez 1992: 285). These text parts are graphically marked: acts and scenes normally start with titles, and utterances, speakers, stage directions and dialogues are graphically delimited. Fig. 1 shows a fragment from *The Crucible* (Miller 1955: 8) where a scene, utterances, speakers, stage directions and dialogues can all be easily identified.

---

[2] TAligner can be accessed at https://addi.ehu.es/handle/10810/42445.

[3] We might mention that, apart from theatre plays, film and TV scripts also belong to the dramatic field (Esslin 1990: 31) and share these structural peculiarities. Therefore, the analysis of these text types could also benefit from the advances in analytical tools suggested here.

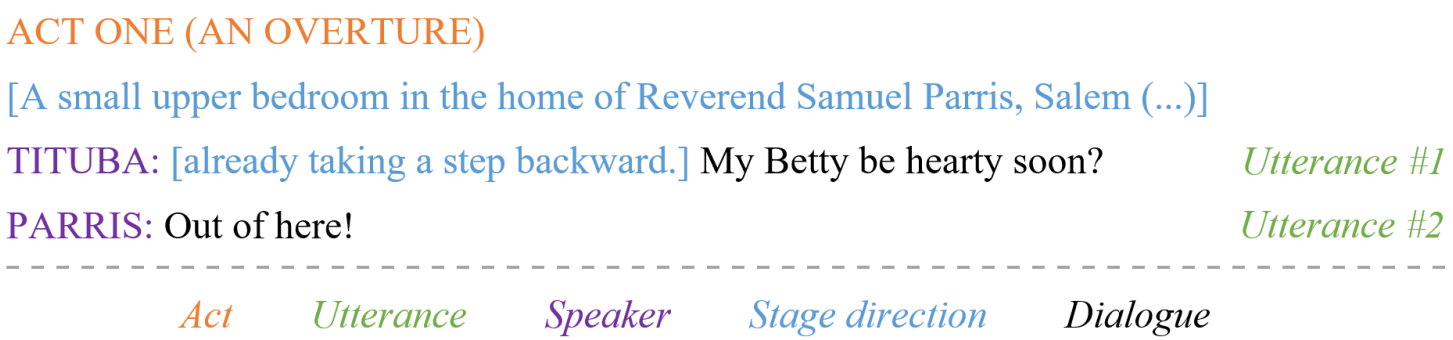

Fig. 1 Structural units of dramatic texts

These structural units of dramatic texts need to be taken into account for both corpus building and analysis. To do this using digital tools, some specific features are required in addition to the functions that are used for dealing with structural aspects common to prose texts. Hence, whereas general software options, such as linguistic annotation, alignment, concordances and statistics continue to be fundamental, three adapted features are desirable in addressing theatrical corpora: the structural annotation of texts, alignment at the utterance level (Merino-Álvarez 1994; Bandín 2007) and analysis options related to the structural annotation. Thus, the specific nature of a text type leads to the need for specific tools, a clear example of the interplay between corpora and software. Table 1 summarises our proposal for general and specific features to build and analyse parallel corpora of theatre plays.[4]

Table 1 Proposed features

Corpus building	Cleaning and validation *Structural annotation of theatre units (TEI, custom XML…) Word-based linguistic annotation *Alignment at the utterance level Texts edition for amendments
Corpus analysis	Introduction of metadata *Concordances filters using theatre structure *Quantitative and qualitative stats based on theatre units

2.1 Structural annotation

The specific units of play texts (acts, scenes, utterances, speakers, stage directions and dialogues) need to be annotated so that they can be processed for alignment and analysis. The implications in terms of software for this type of structural annotation involve establishing a way of recognising the specific units and then employing an annotation system. To facilitate recognition, the text-cleaning phase should standardise signs which delimit units as well as solving issues common to other texts types (such as tokenising, removing running heads, page numbers,

[4] Theatre-specific features are marked with asterisks.

fixing OCR errors, etc.). The detection of units may then take these indicators as a reference for automatic text segmentation. Regarding annotation, there is no consolidated standard for corpora annotation, although different initiatives have been undertaken (Stührenberg 2012). It is worth noting that the "Performance Texts" section of the Text Encoding Initiative guidelines (TEI Consortium 2019: 244-247) includes the structural peculiarities of plays and is useful in terms of replicability. Fig. 2 shows an example of this type of structural annotation.

```
<div1 type="act" n=”1”>
        <head>ACT ONE (AN OVERTURE)</head>
        <stage>A small upper bedroom in the home of Reverend Samuel Parris, Salem
        (...)</stage>
        <sp n="1">
                <speaker>TITUBA:</speaker><stage>already taking a step
                backward</stage> My Betty be hearty soon?
        </sp>
        <sp n="2">
                <speaker>PARRIS:</speaker> Out of here!
        </sp>
</div1>
```

Act *Utterance* *Speaker* *Stage direction* *Dialogue*

Fig. 2 Example of structural annotation

Corpus tools do not typically provide the option of adding structural annotations to build theatre corpora, which has resulted in the need for performing additional and time-consuming tasks. For instance, Culpepper (2014: 13) explains the process that had to be carried out in order to differentiate dialogue from non-speech material and to indicate the character to whom a speech corresponds. This consists of introducing *ad hoc* tags at the beginning and end of each character's utterance (Culpeper 2014: 33). Switch-on tags were added using find and replace operations, but switch-off ones had to be introduced individually (Culpeper 2014: 33). As Culpeper observes (2014: 33), this process should be automatised in the future. Although introducing *ad hoc* tags is a possible solution, it is very time-consuming and may lead to errors, and has to be done from scratch for each play text in a corpus. The creation of scripts to automate structural annotation tasks such as this requires a degree of programming knowledge. In addition, if texts are aligned, the source data cannot be edited to fix errors. For these reasons, in the present paper we argue for an inclusion of structural annotation as part of the features of tool for building a theatre corpus.

It is worth noting that in recent years some research projects have created and made available monolingual corpora of plays that are structurally annotated, which highlights the importance of this issue in theatre corpus compilation. The Encyclopedia of Shakespeare's Language and the Drama Corpora projects are extremely interesting examples here. The former project compiled the *Enhanced Shakespearean Corpus (ESC)*,[5] a valuable resource with a variety of components; among these are the *ESC: First Folio Plus*,[6] a corpus of 36 plays by the British playwright, and *ESC: Comparative Plays*,[7] another corpus including 46 play texts by 24 authors of the same period and genres, to be used as a reference corpus. In these corpora, structural units (acts, scenes, utterances, speakers, stage directions) are annotated with XML. Another corpus with similar structural annotation is the Shakespeare Corpus,[8] containing 37 of his plays. On the other hand, the Drama Corpora project provides a platform that brings together 11 corpora of European plays from different sources (in-house and externally compiled).[9] In these corpora structural units are annotated with XML-TEI. This interest in building theatre corpora seems to focus only on monolingual ones, but it would be desirable to go beyond this and to see a similar growth of interest in parallel corpora. Since this type of corpora requires alignment, the development of tools for structural annotation and alignment will facilitate this.

2.2 Alignment at the utterance level

The utterance has proved to be instrumental as the alignment unit in descriptive-comparative analyses of translated plays (e.g. Merino-Álvarez 1992, 1994, 2007; Pérez 2004; Bandín 2007). Since there were no available tools that allowed this type of alignment, previous alignments had to be performed through tables in text processors, a task that is time-consuming and does not offer the advantages of having specific tools for analysis. However, it seems appropriate to transfer something which worked in previous studies to tools which can be used in new analyses on theatre translations.

On these lines, Bandín (2007: 29) highlights the usefulness of taking the utterance as the alignment unit, and argues for the development of a tool which offers this option:

> the utterance stands as the most adequate minimum unit for comparison and alignment in the case of theatre texts. However, due to the high degree of structural difference which exists between source and target texts, there are not yet any computer programmes available to facilitate their alignment. In general terms, our target texts do not present a formal sentence-to-sentence or paragraph-to-paragraph correspondence with their source text, a criterion which would enable alignment using existing aligning tools such as Translation Corpus Aligner (TCA2). (Bandín 2007: 29, our translation)

---

[5] http://wp.lancs.ac.uk/shakespearelang/project-resources/data/
[6] http://wp.lancs.ac.uk/shakespearelang/files/2019/08/ESC-First-Folio-Plus-Manual32483.pdf
[7] http://wp.lancs.ac.uk/shakespearelang/files/2019/08/ESC-Comparative-Plays-Corpus-Manual32481.pdf
[8] https://lexically.net/wordsmith/support/shakespeare.html
[9] https://dracor.org/

In some cases, as in the one presented here, there may not even be sentence correspondence due to the number of changes in the target texts. Even if there were correspondences in certain translations, alignment at the sentence or paragraph level does not work for this text type. For example, should we attach a speaker's name only to a sentence or to each sentence which comes after that character's name? Also, should we distinguish text segments corresponding to speech and nonspeech or treat them as if they were undifferentiated parts? On the other hand, paragraphs are uncommon in this text type, and thus they do not seem suitable to this end. We share Bandín's view that the utterance is the most adequate unit for alignment. It makes more sense to take into account the inherent structure of play texts and simply reflect this structure when we segment a text for alignment, than to force it into other types of units used for prose in corpus tools.

2.3 Concordances and stats using theatre units

Analysis options should take into account the structural annotation of theatre plays. It is important to be able to filter concordances and statistics, such as keyword lists, according to the structural units, since this enables researchers to focus on the parts of text that they are interested in. Global divisions allow us to choose whether we want to conduct analyses in particular text sections, and finer-grained subdivisions in speakers, stage directions and dialogues allow us to analyse differentiated text levels.

Previous research on theatre corpora shows the need for these analysis filters. For instance, Culpeper (2014: 13) notes the need to distinguish between the dialogue of different characters as well as between dialogue and nonspeech textual material, this in a study of keywords that characterise speakers' interventions in Shakespeare's *Romeo and Juliet*. He uses Wordsmith Tools (Scott 2012), which is not equipped particularly for theatre texts but offers an advanced analysis option that allows one to define "tags to include" and "tags to exclude". In this way, the tags previously added to the corpus units may be inserted there in order to analyse the dialogue of only certain speakers. Even in the best-case scenario of being able to filter searches, the process is more time-consuming and complex than with the adapted features suggested. A further step in analysis options based on the annotation of the structural units of plays is provided for the Enhanced Shakespearean Corpus. This corpus can be accessed via CQPweb at Lancaster[10] and queries can be filtered according to the predefined units, among other aspects. Users need to use CQP syntax, unless they select the restricted mode, which has some predefined options. A different option to access theatre corpora is provided by the Drama Corpora project. The corpora hosted on this platform can be analysed through an API, although this requires some programming knowledge. The platform also offers software called EasyLinavis[11] to create speakers' networks of relations using the structural annotation of speakers

[10] http://corpora.lancs.ac.uk/esc-user-service/
[11] https://ezlinavis.dracor.org/

and acts/scenes. Finally, we might note that these are all examples of monolingual corpora, using tools that do not process parallel corpora, or are platforms where only predefined corpora can be analysed. Thus, there is still a gap in available resources regarding tools that allow users to create and analyse their own parallel (or monolingual) theatre corpora.

3. An overview of TAligner and ACM-theatre

Theatre corpora are somewhat different from other types of corpora in terms of structure. As previously mentioned, play texts are usually organised into different acts and scenes. They are further structured into utterances, which are also the most appropriate alignment unit (Merino-Álvarez 1994: 45; Bandín 2007: 29). Utterances are subdivided into speakers, dialogues and stage directions. An analysis of available software reveals that most applications lack any type of specific support for theatre texts in this sense (Scott 2012; Anthony 2014; Kilgarriff et al. 2014). Standard corpus analysis tools, aimed at being useful in the most common textual scenarios, are too general to be used for an exhaustive corpus analysis of play texts. First, they do not annotate the units of theatre texts such as utterances, speakers, stage directions, dialogues, acts or scenes, and hence a differentiation of such parts of the text is not maintained and cannot be used for alignment or analysis. Since utterances are not used as alignment units, it is not possible to relate speakers and their speech or stage directions. Sentences or paragraphs also fail to meet the alignment needs of a type of translation characterised by variability (Merino-Álvarez 1994: 43; Bandín 2007: 29). Moreover, lack of support for analysis options that use the annotation of the structural units of plays hinders studies of a text type characterised by the presence of distinct language levels, basically speech and nonspeech material (Merino-Álvarez 1994: 44). Some tools offer complicated ways to search areas of XML, such as the advanced tags in WordSmith Tools. A tool that enables users to search among utterances, stage directions, global divisions and characters, would be of great benefit to researchers.

Furthermore, as Sanjurjo-González (2018: 47-48) observes, there are fewer software options that process parallel corpora than monolingual ones. The most common ones are AntPConc (Anthony 2014), a parallel version of AntConc, Sketch Engine (Kilgarriff et al. 2014), and CQPweb (Hardie 2012). While all of these allow for building parallel corpora in one way or another (Sanjurjo-González 2018: 31, 34, 38), none includes any built-in aligner.[12] Even if an external aligner is used and output texts are formatted following the corpus analysis software requirements (a process which may demand some technical knowledge), sentences or paragraphs are used as

[12] Evert (2014) points out that “cwb-align isn't a particularly sophisticated sentence aligner, so it's likely to get some cases wrong”.

alignment units instead of utterances.[13] Taken together, all this makes it very hard to handle theatre corpora using general tools such as these unless users have some programming skills.

As a consequence of this lack of software, most parallel corpora of theatre translations (Merino-Álvarez 2007; Bandín 2007; Pérez 2004) have been aligned through repetitive and time-consuming methods, for instance using tables and cells, and then studied using the functions of text processors. Apart from the laborious work involved, analysis options for these corpora are also limited when compared to analysis using specific tools.

Since theatre plays mirror spoken discourse, they both include an organisation into utterances linked to speakers. This partial affinity of theatre scripts with the type of discourse they imitate has also led us to consider whether it might be useful and straightforward to apply tools used for spoken corpora to theatre corpora. A brief analysis of such tools reveals that they are not wholly suitable for building and performing linguistic analyses of theatre corpora, since they do not include all the required features (most only mark speakers as units), and their use would involve a complex process in which different software would need to be used for annotation and querying. In addition, query tools present complication in terms of the multiple types of annotation that they can handle. For instance, as Zeldes et al. (2009: 358) note, query results can be exported in order to process them with data mining tools; however, if a user chooses to employ spoken corpus-related tools, they must be aware that manual annotation needs to be performed both for units and for alignment, and subsequently a compatible (and not theatre-specialised) software will have to be used for querying the corpus. Thus, it requires more time and complex operations than our proposal for integrating all the required features in a single tool.

TAligner and ACM-theatre offer new options for the specific building and analysis theatre corpora. Both tools overcome the main issues of general corpus tools: they recognise the structure of theatre plays, that is, the presence of utterances, speakers, stage directions and dialogues, among others, and they also make it possible to align and analyse parallel corpus using those specific units.

3.1 TAligner

TAligner is a corpus analysis tool that enables users to build parallel corpora of theatre, narrative and poetry texts, as well as querying them through an intuitive process. Within the framework of the TRACE projects, based in Spain (University of León and University of the Basque Country), the adaptation of a corpus software for theatre texts was felt necessary and was implemented for a first version of the application: TRACE Corpus Tagger/Aligner 1.0 (Gutiérrez-Lanza, Bandín, García-González and Lobejón-Santos 2015). The development of this tool was possible thanks to the collaboration between researchers from the TRALIMA/ITZULIK research group and

[13] Line breaks might be useful for alignment; however, utterances can have more than one paragraph. In addition, the use of line breaks in plays may be inconsistent. Moreover, features of AntPConc are far from those of the tool's monolingual version. AntPConc can be considered as a simple parallel concordance.

computer scientist Iñaki Albisua, and led to the current version, TAligner 3.0 (Sanz-Villar and Andaluz-Pinedo 2021).

Strictly speaking, TAligner is a 3rd generation corpus tool (McEnery and Hardie 2012: 37-48). As such, it does not use a client/side paradigm, which undoubtedly affects its accessibility. It has been developed using the Java programming language, so it is necessary to install a Java Virtual Machine to run the program. Regarding the user interface, it offers a very simple and intuitive menu that enables users to clean, annotate, align, edit and query their corpora. It does not use any advanced search system, so it merely transforms users' input into equivalent regular expressions through a drop-down menu.

In order to use the part of TAligner designed for theatre corpora, the theatre option simply needs to be selected at the start from the left-hand side menu, as shown in Fig. 3. Theatre corpus building and analysis processes may then be carried out using the horizontal menu.

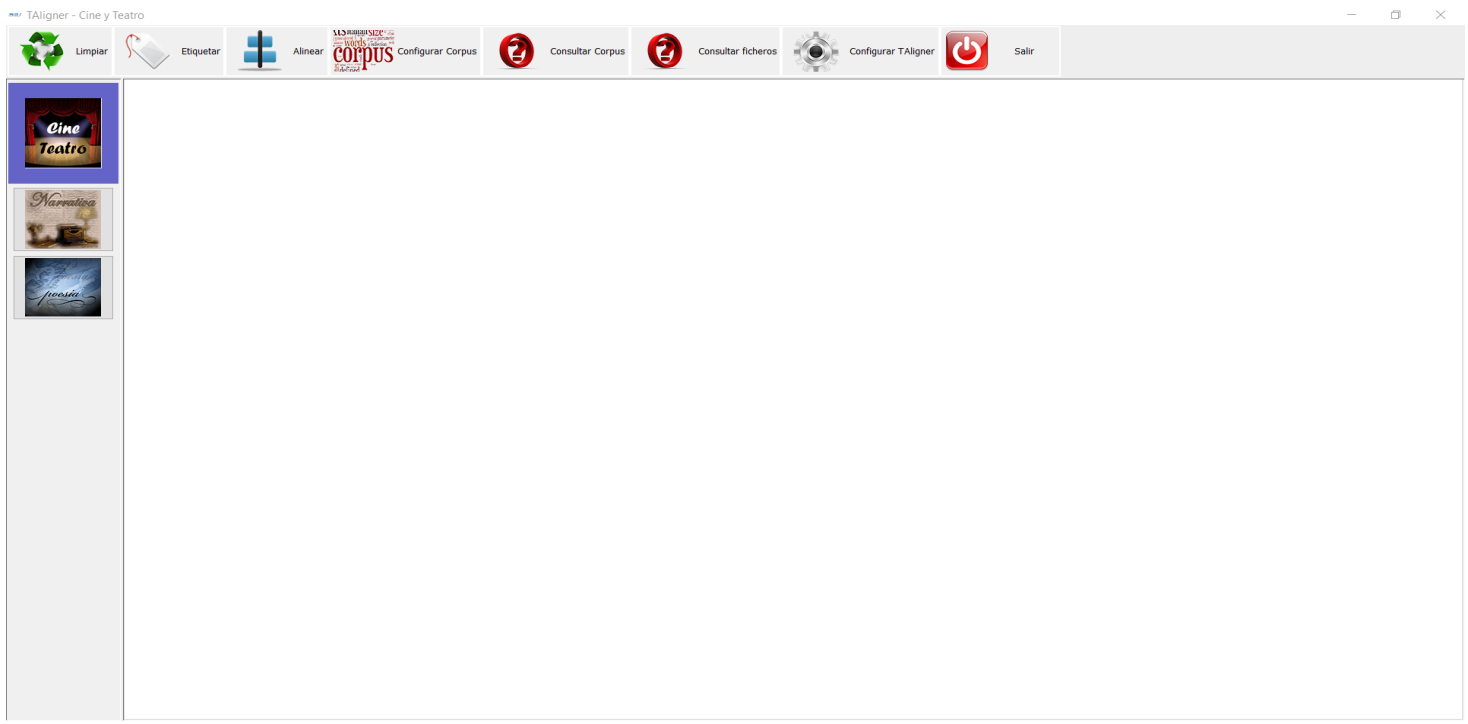

Fig. 3 Selection of TAligner's theatre section

For corpus building, automatic cleaning can first be performed (Fig. 4). Apart from recurrent errors common to narrative texts such as double spaces, the application standardises the punctuation marks which signal the end of character names and delimit stage directions to facilitate subsequent recognition of those units.

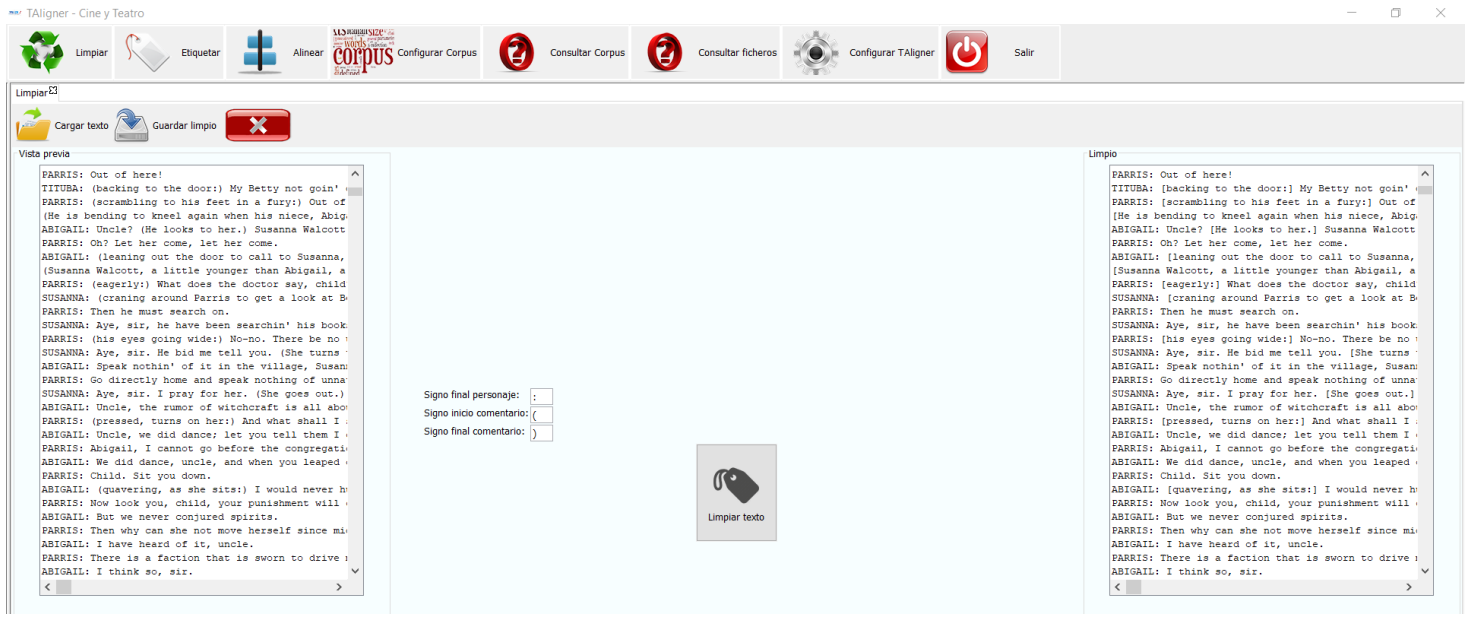

Fig. 4 Data cleaning in TAligner

In the following tab, structural annotation is introduced automatically to utterances, as well as speakers, stage directions, dialogue, acts and scenes, as shown in Fig. 5.

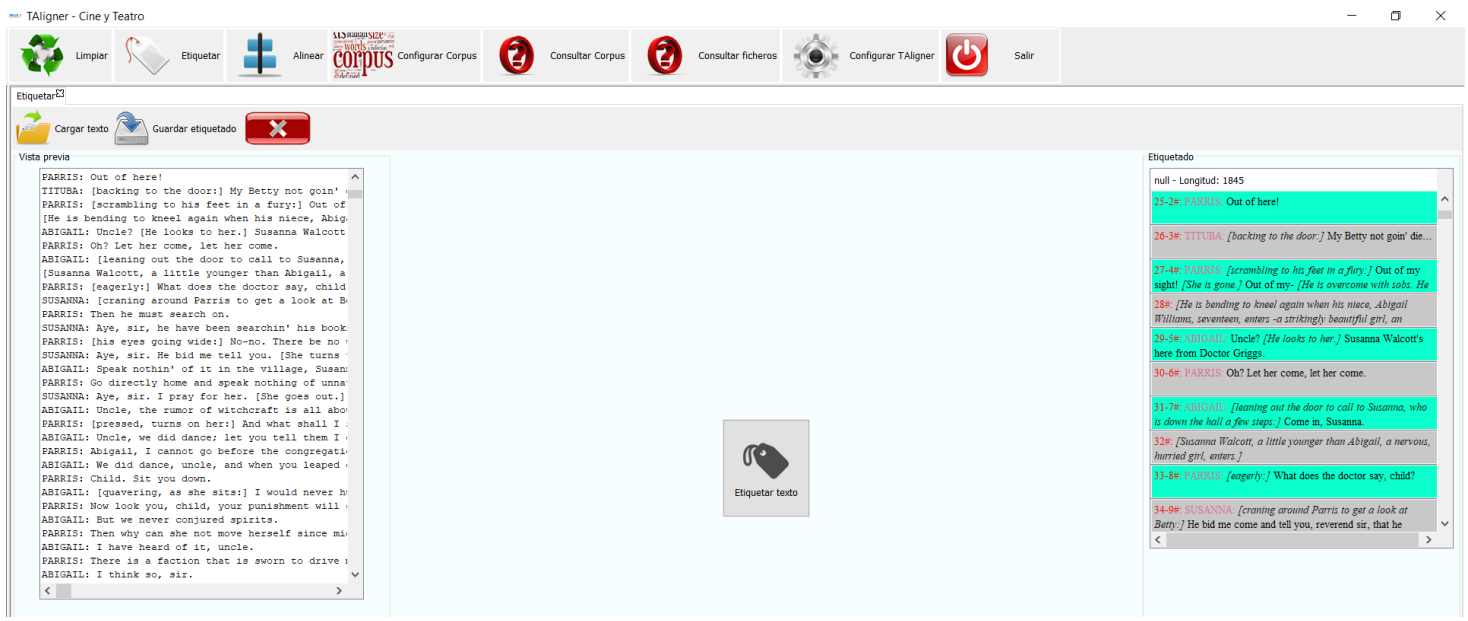

Fig. 5 Structural annotation in TAligner

TAligner offers a built-in aligner that takes the utterance as the alignment unit. More than two texts may be aligned simultaneously, a useful option when dealing with retranslations or intermediary texts. The aligning task is performed manually. This might be regarded as an inconvenience or an advantage depending on the case, since it is time-consuming but also reduces the alignment error rate. The tool provides an intuitive user interface to carry

out this action (Fig. 6). It includes options such as merging or splitting utterances, inserting a blank segment, or even deleting one. Another useful option for corpus building within the tool is editing, which allows users to easily modify both the text and the structural annotation in case an error is detected at this stage. If necessary, editing also allows one to introduce observations by way of a custom annotation system that can be preestablished in the tool's settings. Apart from these features, in this screen users can associate metadata to texts such as author, title and translator. Once the aligned texts are ready, they are simply added to a corpus within the tool, and it is established which texts are the source and target ones. Alignments as well as individual aligned texts can also be saved locally as TMX or XML files, respectively. A possible drawback in terms of the tool's usability is that it only processes texts with the annotation that it produces, since the software is designed to carry out the whole process.

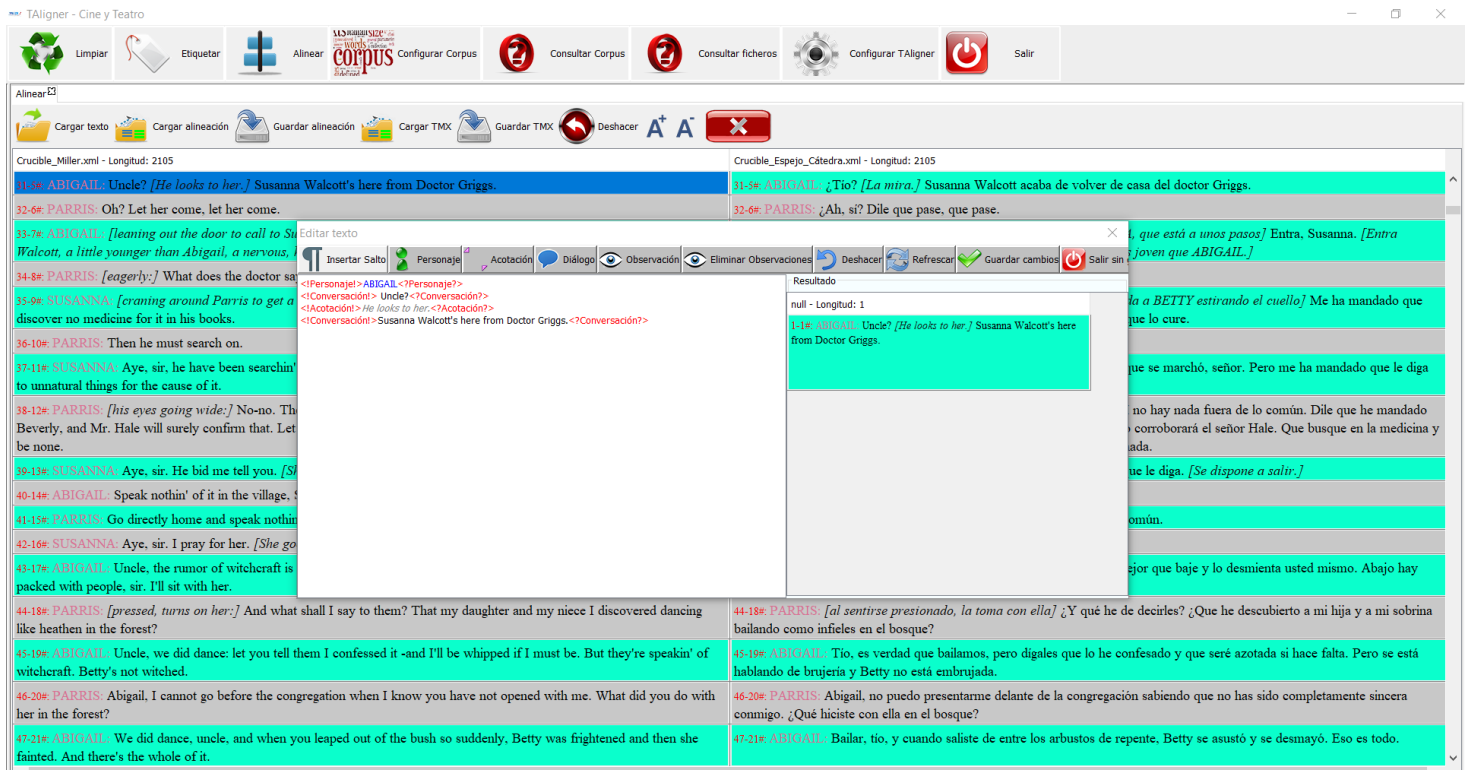

Fig. 6 Aligning screen in TAligner

Regarding analysis features, TAligner only offers concordances (Fig. 7). The interface includes a dropdown menu which allows users to choose whether they are interested in searching whole texts, only dialogue, or only stage directions. For instance, if we want to examine how the discourse marker *well* has been translated in a parallel theatre corpus, we focus only on dialogue to avoid occurrences of this word in stage directions. Although results still include some different functions of *well* in dialogues, a great deal of noise in the data is avoided through this operation. Although this filter does not include speakers, acts or scenes, it would also be interesting to add these

units in the future as they are also annotated within the tool. One issue of this tool, compared to ACM-theatre, is that it does not include specific statistics for theatre corpora.[14]

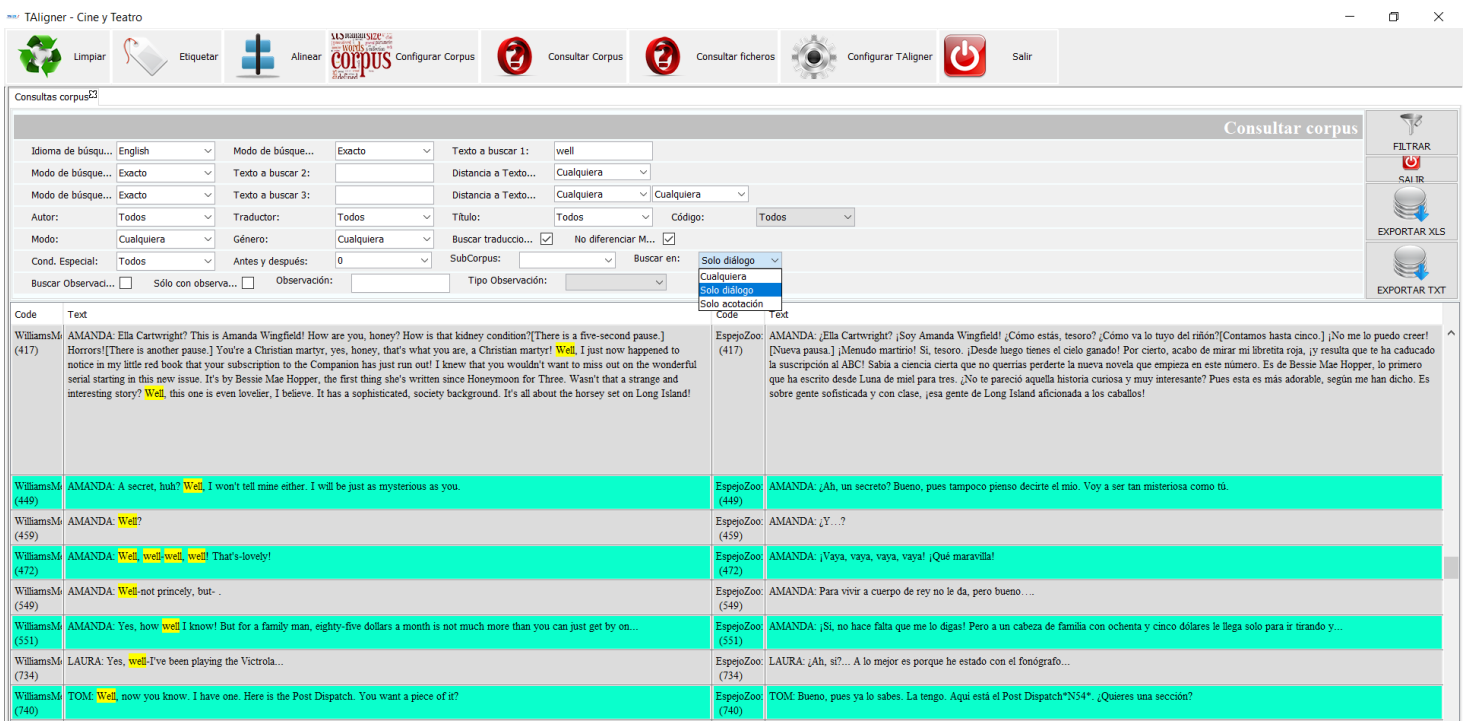

Fig. 7 Query section in TAligner

In sum, TAligner allows for the creation of parallel theatre corpora that are structurally annotated (acts/scenes, utterances, speakers, stage directions and dialogues) and aligned at the utterance level. Corpus building in this tool provides some additional features, such as text editing for amendments, alignment of multiple retranslations to the same original text, custom annotation, and the addition of metadata. In relation to analysis, filters for stage directions and dialogue add to a range of query options. While there are issues that could be improved in the future in relation to analysis statistics and linguistic annotation, the tool as developed thus far has made it possible to create parallel corpora of theatre translations (Merino-Álvarez and Andaluz-Pinedo 2017; Sanz-Villar and Andaluz-Pinedo 2021) that add to the narrative corpora already compiled using this software (Arrula 2018; Sanz-Villar 2015, 2019; Zubillaga 2013; Zubillaga, Sanz-Villar and Uribarri 2015).

3.2 ACM-theatre

[14] The application offers frequency lists, but so far they take texts as wholes (Sanz-Villar and Andaluz-Pinedo 2021).

ACM-theatre[15] (Sanjurjo-González 2017a, 2017b, 2018) was developed as part of a doctoral dissertation within the ACTRES research group. It was originally developed as software for corpus linguistic analysis that allows users to build bi/multilingual, comparable and monolingual corpora without technical assistance, and be able to annotate, align and process these at different layers. Support for theatre corpora was adopted during a further collaboration between ACTRES and TRALIMA/ITZULIK research groups, based on users' experience with TAligner to build and analyse theatre corpora. The development of this tool has led to an increase in the possibilities for compilation and analysis of parallel theatre corpora.

From a technical point of view, ACM-theatre is a 4th generation concordancer that allows users to access and query their corpora from any device with internet connectivity and a web browser. It has been developed using CWB (Evert and Hardie, 2011) as a back-end. A visual query system using selectors based on P-ACTRES 2.0 (Sanjurjo-González and Izquierdo, 2019: 224-226) is linked to the CQP query language (Evert, 2020) and allows users to make complex queries including regular expressions, linguistic annotations, characters or specific parts of a theatre play, or even simultaneous queries over the different subcorpora, all without any technical knowledge of the query language.

In order to build a theatre corpus, users select files and the linguistic annotations they want to add (Fig. 7).

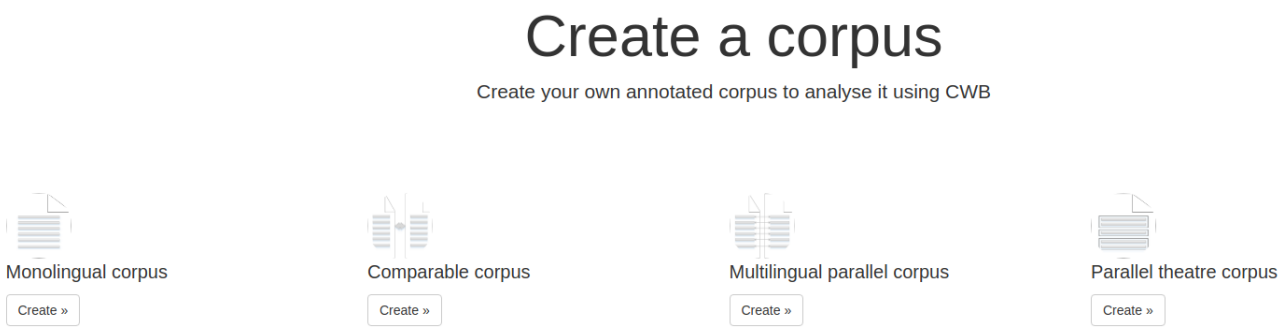

Fig. 8 Corpus building screen in ACM-theatre

Before texts are uploaded to the corpus an adapted cleaning process based on the most common errors in play texts is used. Table 3 shows some of these errors.

Table 2 Some errors in OCR scanned theatre scripts

Problem	Before cleaning	After cleaning
Speaker names are standardised as ending with a colon.	***ABIGAIL*** *(He looks to her.)**:** Susanna Walcott's here from Doctor Griggs.*	***ABIGAIL:*** *(He looks to her.) Susanna Walcott's here from Doctor Griggs.*

[15] ACTRES Corpus Manager pending register.

Each utterance starts in a new line.	*ABIGAIL: (He looks to her.) Susanna Walcott's here from Doctor Griggs.* ***PARRIS: Oh? Let her come, let her come.***	*ABIGAIL: (He looks to her.) Susanna Walcott's here from Doctor Griggs.* ***PARRIS: Oh? Let her come, let her come.***
Extra line breaks are removed.	*PARRIS: Oh? Let her come, let her* ¶ *come.*	*PARRIS: Oh? Let her come, let her come.* ¶

Following this, ACM-theatre automatically annotates and aligns different sections of the corpora (Fig. 9). As previously mentioned, play texts have a very specific document format in which speakers generally appear in capitals, preceded by a line break and followed by a colon, and hence they are easy to identify. The aligning process takes this name as a reference in order to identify the utterances that are later used as alignment units. Regarding the rest of the components, stage directions are identified as the text parts between brackets and dialogue appears unmarked. On the other hand, acts and scenes are detected through their titles. However, this approach may provide erroneous results in the alignment process of some documents and in noisy corpora. Noisy corpora of this text type may be related to language transfer in which there are many utterance additions or omissions, for instance in the case of stage adaptations. For all these reasons, automatic alignment will be more suitable for theatre translations which do not involve significant structural changes.

```
<s speaker="ABIGAIL">
Uncle?
<stage>
He          he      PP      Z8m
looks       look    VVZ     S7.1-[i1.2.1 A8 X3.4 X2.4 X7+ Z4
to          to      TO      S7.1-[i1.2.2 Z5
her         her     PP      Z8f
.           .       SENT
</stage>
Susanna     Susanna NP      Z1mf
Walcott     Walcott NP      Z1mf
's          's      POS     A3+ Z5
here        here    RB      M6 T1.1.2 T1.2
from        from    IN      Z5
Doctor      Doctor  NN      B3/I3.2/S2mf P1/S2mf B3 A2.1+/G2.1-
Griggs      Griggs  NP      Z99
.           .       SENT
</s>
```

Fig. 9 Internal document format in ACM-theatre with lemma, POS and semantic information

Regarding analysis features, ACM-theatre provides all the common corpus linguistic analysis options, such as collocates, keywords, frequencies lists, n-grams extraction, among others. It includes some linguistic annotations at the word level, such as grammatical and semantic ones. It also provides some specific theatre-related features, such as the ability to filter queries according to the text divisions into speakers, stage directions, dialogues, acts

and scenes. Furthermore, these units could be used for statistics (Fig. 10). An example of a query is shown in Fig. 11.

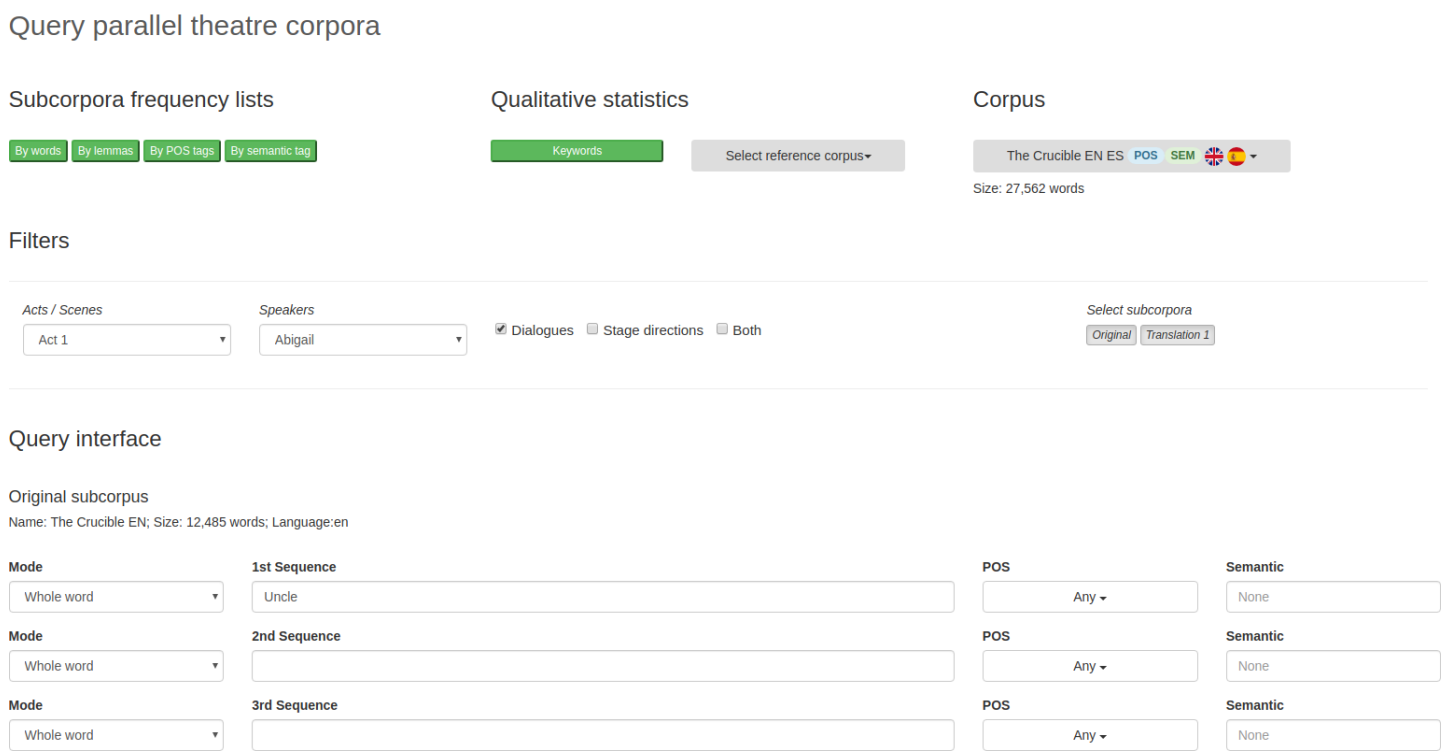

Fig. 10 Some analysis options in ACM-theatre

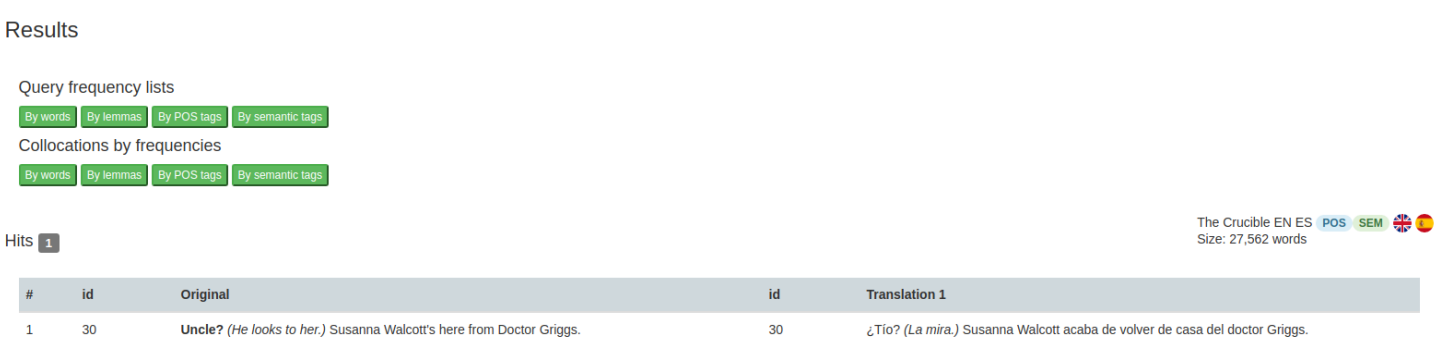

Fig. 11 Query result in ACM-theatre

All in all, ACM-theatre offers the user a wide range of options to create and analyse corpora without the need for technical assistance (Sanjurjo-González 2018: 99). Corpus building adds structural annotation to differentiate the units of plays (acts, scenes, utterances, speakers, stage directions and dialogue) and aligns texts automatically at the utterance level. ACM-theatre does not use any annotation standard, so it is not compatible with TEI performance guidelines. It allows users to compare up to four different texts but it does not include any metadata or the option to edit texts and tags in the case a mistake is spotted. This tool also provides grammatical and semantic annotation.

Regarding analysis, it offers concordances and other common options (such as collocates, keywords, frequencies lists, n-grams extraction) that use the specific units of plays.

3.3 Comparison of TAligner and ACM-theatre

Both TAligner and ACM-theatre allow users to build and analyse parallel theatre corpora, since they annotate the structural units of plays, and use that annotation to align texts at the utterance level, as well as filter analyses. Nevertheless, the two tool offers complementing approaches to building and querying corpora, and may be found to be suitable for different purposes. Regarding corpus building, the tools differ mostly in their options for alignment and linguistic annotation. TAligner offers the possibility of editing alignments, which is helpful if the user needs to achieve the highest precision rate possible, whereas ACM-theatre opts for automatic alignment, which reduces considerably the required time for the task but at the expense of a lower precision rate. On the other hand, TAligner includes the possibility of adding custom annotation, although other types of linguistic annotation are not yet implemented. Within ACM-theatre, corpora are annotated grammatically with POS tags, using SpaCy[16], and semantically, according to the USAS Category System (Archer, Wilson y Rayson 2002).[17] In relation to corpus analysis, two essential differences are also found between the tools in terms of the techniques available and the type of elements that may be queried. First, TAligner makes it possible to retrieve concordances, but it does not include statistics that take into account the structural annotation of play texts. However, ACM-theatre, apart from concordances, offers common statistics such as keyword lists. Second, the analysis options linked to linguistic annotation logically differ in the tools: TAligner allows the searching of custom annotations, and ACM-theatre the searching of grammatical and semantic annotations.

All in all, TAligner and ACM-theatre are both useful for processing parallel corpora of plays but they include different functionalities. Users might use TAligner or ACM-theatre based on their particular needs for corpus building and analysis. If precise alignment is needed and the corpus analysis relies solely on concordances and/or custom annotation, TAligner is an appropriate choice. In addition, its ease of use makes it appropriate not only for more technically proficient researchers but also for university students (Sanz-Villar and Andaluz-Pinedo 2021). On the other hand, ACM-theatre offers more advanced analysis options (such as keywords or collocations) which are extremely valuable for specialists in linguistics and translation. Table 3 provides a summary of the main features available for building and analysing parallel theatre corpora in TAligner and ACM-theatre.

---

[16] https://doi.org/10.5281/zenodo.1212303

[17] A recently developed tool for custom annotation, OpenTagger (Sanjurjo-González and Andaluz-Pinedo 2020), is planned to be integrated into ACM-theatre in the future.

Table 3 Summary of features for building and analysing parallel theatre corpora in TAligner and ACM-theatre

Tasks	Features	TAligner	ACM-theatre
Corpus building	Cleaning and validation	✓	✓
	*Structural annotation of theatre units	XML	XML
	Word-based linguistic annotation	✗	✓
	*Alignment at utterance level	✓	✓
	Texts edition for amendments	✓	✗
Corpus analysis	Introduction of metadata	✓	✗
	*Concordances filters using theatre structure	✓	✓
	*Quantitative and qualitative stats based on theatre units	✗	✓

4. Conclusions

New advances in parallel corpora tools have made it possible to accommodate specific structural features of theatrical texts. The focus of this contribution is on parallel corpus building and analysis software, and how it can deal with this text type. Since corpus tools play a fundamental role in corpus-based research, this new direction is essential for studies on parallel corpora compiled from theatre translations. Our aim is to shed some light on the scope of TAligner and ACM-theatre to create and analyse parallel theatre corpora, and thus to contribute to progress on these lines.

Theatre plays possess an inherent basic structure in terms of acts, scenes, utterances, speakers, stage directions and dialogues. Based on this, at least three functionalities are needed for the construction and analysis of parallel corpora of play texts: structural annotation, alignment at the utterance level, and concordances and statistics using these units. A review of general tools for corpus building and analysis shows that these features are missing, or that complex and time-consuming operations need to be carried out. The situation is even more critical for parallel corpora, since there are far fewer options that allow for working with aligned corpora.

In this paper we introduce two applications that were recently designed to deal specifically with theatre texts, TAligner and ACM-theatre. TAligner allows for marking up the units, manually aligning texts at the utterance level, and searching the resulting theatre corpora according to different units. ACM-theatre also recognises the structure of theatre texts, aligns automatically these texts at the utterance level, and analyses units through searches and statistics. The different characteristics of these two applications for use with prose texts with regard to corpus building and analysis are reflected in their versions for theatre texts. TAligner involves manual alignment, has the option of introducing custom annotations, and most analysis options are related to searches. On the other hand, ACM-theatre provides automatic alignment, adds grammatical and semantic annotation, and offers the possibility of searching and extracting a wide range of statistics, including keyword lists from each subunit. These tools, then, provide solutions for different needs in the process of making and using parallel corpora of theatre translations.

Looking at the future, the adaptation of tools for theatre texts opens the way to other developments which may be found useful in further studies. For instance, the software discussed here may be improved, including the

incorporation of features which are absent from the current versions. For instance, TAligner could add a linguistic tagger or a basic statistics package; ACM-theatre might include an interface for editing erroneous alignments. In both tools it would be beneficial to support other types of document formats such as TEI (TEI Consortium 2019: 244-247), so that users could use previously compiled corpora of play texts in that format. Another means of improvement would be through new advances in the software options for parallel corpora in general, with these advances transferred to theatre texts. In this sense, word alignment would increase analysis options for contrastive and translation studies. Finally, further developments could be oriented towards multimodal corpora by means of audio and video recordings of plays. More studies are needed and further software improvements could be implemented here. However, even as we currently stand, the development of parallel corpus software for play texts has begun to bridge the gap in available resources, and will undoubtedly contribute to future research on theatre translations.

Acknowledgments

Research group TRALIMA/ITZULIK, GIU 16/48, University of the Basque Country, UPV/EHU, Basque Government consolidated research group IT1209-19.

Research group ACTRES. Part of this study has been supported by the Spanish Agency for Research, Development and Innovation (Ministry of Economy and Competitiveness) [FFI2016-75672-R].

Red de Excelencia CorpusNet, funded by Ministry of Economy and Competitiveness project [FFI2016-81934-RED].

At the time of writing, the co-author Olaia Andaluz-Pinedo is a doctoral student funded by the University of the Basque Country UPV/EHU, Spain.

We would like to thank the reviewers for their useful comments.

References

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research,* https://doi.org/10.17250/khisli.30.2.201308.001

Anthony, L. (2014). AntPConc (Versión 1.1.0) [Software]. Tokio: Waseda University. http://www.laurenceanthony.net/ Accessed 3 April 2020

Archer, D., Wilson, A. & Paul Rayson (2002). Introduction to the USAS category system. *Benedict project report*. http://ucrel.lancs.ac.uk/usas/usas%20guide.pdf

Arrula, G. (2018). *Autoitzulpenaren teoria eta praktika Euskal Herrian / Theory and practice of self-translation in the Basque Country* (Doctoral dissertation). Bilbao: Universidad del País Vasco. http://hdl.handle.net/10810/27983 Accessed 3 April 2020

Bandín, E. (2007) *Traducción, recepción y censura de teatro clásico inglés en la España de Franco. Estudio descriptivo-comparativo del Corpus TRACEtci (1939-1985)* (Doctoral dissertation). León: Universidad de León. https://buleria.unileon.es/handle/10612/1885 Accessed 3 April 2020

Culpeper, J. (2014). Keywords and Characterization. An Analysis of Six Characters in *Romeo and Juliet.* In D. L. Hoover, J. Culpeper, & K. O'Halloran (Eds.), *Digital Literary Studies. Corpus Approaches to Poetry, Prose and Drama* (pp. 9-33). New York: Routledge.

Doval, I., & Sánchez-Nieto, T. (2019b). Parallel corpora in focus: An account of current achievements and challenges. In I. Doval Reixa, & M. T. Sánchez Nieto (Eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

Esslin, M. (1990). *The Field of Drama*. London: Methuen.

Evert, S. (2014). [CWB] A question about the aligning using cwb-encoding (CWB mailing list). http://liste.sslmit.unibo.it/pipermail/cwb/2014-January/001529.html Accessed 3 April 2020

Evert, S. (2020). The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial http://cwb.sourceforge.net/files/CQP_Tutorial.pdf Accessed 2 September 2020

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics Conference 2011. Birmingham: University of Birmingham.

Gutiérrez-Lanza, C., Bandín, E., García-González, J. E., & Lobejón-Santos, S. (2015). *Desarrollo de software de etiquetado y alineación textual: TRACE Corpus Tagger/Aligner 1.0©.* Paper presented at the II Congreso Internacional de Humanidades Digitales Hispánicas: Innovación, globalización e impacto, Madrid, Spain. http://hdh2015.linhd.es/ebook/hdh15-gutierrezlanza.xhtml Accessed 3 April 2020

Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics,* https://doi.org/10.1075/ijcl.17.3.04har

Johansson, S., & Hofland, K. (1994). Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie, & P. Schneider (Eds.), *Creating and Using English Language Corpora*, (pp- 25-37). Amsterdam: Rodopi.

Kenny, D. (2001). *Lexis and creativity in translation: a corpus-based study*. Manchester: St. Jerome.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography,* https://doi.org/10.1007/s40607-014-0009-9

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X*(5), 79-86.

Lavid, J. (2019). Discourse annotation in the MULTINOT corpus: Issues and challenges. In I. Doval Reixa, & M. T. Sánchez Nieto (Eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

Marco, J. (2019). Living with parallel corpora: The potentials and limitations of their use in translation research. In I. Doval Reixa, & M. T. Sánchez Nieto (Eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

McEnery, T., & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Merino-Álvarez, R. (1992). Rewriting for the Spanish stage. *KOINÉ. Annali della Scuola Superiore per Interpreti e Traduttori San Pellegrino, 2*(1-2), 283-289.

Merino-Álvarez, R. (1994). *Traducción, tradición y manipulación. Teatro inglés en España 1950-1990*. León: Universidad de León / Bilbao: Universidad del País Vasco.

Merino-Álvarez, R. (2007). La homosexualidad censurada: estudio sobre corpus de teatro TRACEti inglés-español (desde 1960). R. Merino-Álvarez (ed.) *Traducción y censura en España (1939-1985). Estudios sobre corpus TRACE: cine, narrativa, teatro*. León: Universidad de León / Lejona: Universidad del País Vasco.

Merino-Álvarez, R., & Andaluz-Pinedo, O. (2017). Peter Shaffer en la cultura española. *Creneida. Anuario de Literaturas Hispánicas*, (5), 239–278.

Miller, A. (1955). *The Crucible*. New York: Penguin Books.

Molés-Casés T., & Oster U. (2019). Indexation and analysis of a parallel corpus using CQPweb: The COVALT PAR_ES Corpus (EN/FR/DE > ES). I. Doval Reixa, & M. T. Sánchez Nieto (Eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

Oksefjell, S. (1999) A Description of the English-Norwegian Parallel Corpus. *International Journal of Corpus Linguistics, 4*(2), 197-219.

Pérez, M. (2004). *Traducciones censuradas de teatro norteamericano en la España de Franco (1939-1963)* (Doctoral dissertation). Bilbao: Universidad del País Vasco.

Rafalovitch, A., & Dale, R. (2009). United Nations general assembly resolutions: A six-language parallel corpus. In *MT Summit XII*, (pp. 292–299). Ottawa: AMTA.

Sanjurjo-González, H. (2017a). *ACTRES Corpus Manager*. [Computer software].

Sanjurjo-González, H. (2017b). *Creación de un framework para el tratamiento de corpus lingüísticos - Development of a framework for corpus linguistic* (Doctoral dissertation). Universidad de León, León, Spain. https://buleria.unileon.es/handle/10612/6920 Accessed 3 April 2020

Sanjurjo-González, H. (2018). *Creación de un framework para el tratamiento de corpus lingüísticos (Development of a framework for corpus linguistic análisis)*. León: Universidad de León, Área de Publicaciones.

Sanjurjo-González, H., & Izquierdo, M. (2019). P-ACTRES 2.0: a parallel corpus for cross-linguistic research. I. Doval Reixa, & M. T. Sánchez Nieto (eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

Sanjurjo-González, H., & Andaluz-Pinedo, O. (2020). OpenTagger: A flexible and user-friendly linguistic tagger. *56th Linguistics Colloquium*. http://hdl.handle.net/10810/48683

Sanz-Villar, Z. (2015). *Unitate fraseologikoen itzulpena: alemana-euskara. Literatur testuen corpusean oinarritutako analisia* (Doctoral dissertation). University of the Basque Country UPV/EHU. http://hdl.handle.net/10810/15128 Accessed 3 April 2020

Sanz-Villar, Z. (2019). An Overview of Basque Corpora and the Extraction of Certain Multi-Word Expressions from a Translational Corpus. I. Doval Reixa, & M. T. Sánchez Nieto (eds.). *Parallel corpora for contrastive and translation studies: New resources and applications.* Amsterdam/Philadelphia: John Benjamins.

Sanz-Villar, Z., & Andaluz-Pinedo, O. (2021). TAligner 3.0: a tool to create parallel and multilingual corpora. J. Lavid, C. Maíz-Arévalo, & J. R. Zamorano-Mansilla (Eds.) *Corpora in translation research: recent advances and applications* (126-146). Amsterdam / Philadelphia: John Benjamins.

Scott, M. (2012). WordSmith Tools (Versión 6) [Software]. Stroud: Lexical Analysis Software. Retrieved from http://www.lexically.net/wordsmith/ Accessed 3 April 2020

Stührenberg, M. (2012). The TEI and current standards for structuring linguistic data. An overview. *Journal of the Text Encoding Initiative*, (3).

TEI Consortium (2019). Performance Texts. *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (pp. 234-259). https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

Zubillaga, N. (2013). *Alemanetik euskaratutako haur- eta gazte-literatura: zuzeneko nahiz zeharkako itzulpenen azterketa corpus baten bidez* (Doctoral dissertation). Bilbao: Universidad del País Vasco. http://hdl.handle.net/10810/12431 Accessed 3 April 2020.

Zubillaga, N., Sanz-Villar, Z., & Uribarri, I. (2015). Building a trilingual parallel corpus to analyse literary translations from German into Basque. In C. Fantinuoli, & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 71–92). Berlin: Language Science Press.