Coreference Annotator
A new annotation tool for aligned bilingual corpora

Mara Tsoumari
School of English
Faculty of Philosophy
Aristotle University of Thessaloniki
54 124, P.O. Box 58, Thessaloniki, Greece
mtsoum2@gmail.com, mara@optimum-services.com

Georgios Petasis
Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications
National Centre for Scientific Research (N.C.S.R.) "Demokritos"
GR-153 10, P.O. BOX 60228, Aghia Paraskevi, Athens, Greece
petasis@iit.demokritos.gr

Abstract

This paper presents the main features of an annotation tool, the Coreference Annotator, which manages bilingual corpora consisting of aligned texts that can be grouped in collections and subcollections according to their topics and discourse. The tool allows the manual annotation of certain linguistic items in the source text and their translation equivalent in the target text, by entering useful information about these items based on their context.

1 Introduction

The annotation tool, Coreference Annotator, has been developed within the framework of wider research in the analysis of parallel texts from a translation point of view. More specifically, the research attempts a theoretical classification of the translation of European Union texts in the light of Relevance Theory (Tsoumari, 2008), and examines a special use monodirectional bilingual corpus consisting of aligned English (originals/source texts) and Greek (translations/target texts) versions of press releases of the European Commission.

The aim of the annotation tool is for the researcher to trace and annotate manually certain linguistic items in the source text and their translation equivalent in the target text, by entering useful information about these items based on their context. The focus for this study is on identifying discourse markers and conjunctions that express concession/contrast/adversity in the source text and then locating their translation equivalent in the target text. To the group of markers mentioned above, the conjunction 'and' has been added. Cases of omission of source text conjunctions or discourse markers, or addition of conjunctions or discourse markers in the target text are also marked.

2 Motivation

The scope of the research that motivated the creation of this tool combines mainly translation, parallel corpora (original-source texts and translation-target texts), semantics, pragmatics, and discourse. A parallel aligned corpus of press releases of the European Commission is examined both translationally and linguistically to reach conclusions about how certain linguistic items are translated, potentially reflecting the intention of the authors; the expectations of the readers; whether intentionality and expectations change when moving from the source text to the target text; and effects from genre, discourses depending on the topics of the documents, public sentiment or culture.

2.1 Translation in the EU

There is an intriguing matter in the translation of European Union documents into all or some of the official languages of the European Union. On the one hand, there are rules and regulations governing the operation of European countries together as a whole, as a single unity forming the European Union, and EU culture and mentality. On the other hand, the European countries-member states maintain their national cultures and mentalities. Research has shown that the culture of the

Proceedings of Recent Advances in Natural Language Processing, pages 43–52,
Hissar, Bulgaria, 12-14 September 2011.

EU edifice is different from national cultures, has a culture of its own, despite the likely blurred borderlines between them (Koskinen, 2001; Koskinen, 2004). EU texts and their translation serve a primary communicative situation, since original texts are written to be translated so as to help EU (source text) authors reach different national (target) language users. Some of the characteristics of EU texts are that they are often produced and translated almost at the same time (Koutsivitis, 1994); translation may constitute the starting point to improve the 'original' (Koutsivitis, 2003); the writers are usually a group of people or a committee; most source texts are written in English and to a lesser degree in French and German (three procedural languages); the authors are not necessarily native speakers of the language they use for writing; source texts may not always be written in one language and have special linguistic, syntactical and stylistic characteristics called Eurojargon, Eurobabble or Eurospeak (Trosborg, 1997). Thus the translation process, strategies and methods are also affected by the particular circumstances of the production of target texts.

2.2 Press releases of the European Commission

EU press releases are one of the types of documents produced in the framework of the European Union and are distinct from non-EU press releases. The reason is that if we accept that the European Union has a culture of its own, as Koskinen (2001; Koskinen (2004) argues, then it is only normal to expect the production of EU culture-specific texts and genres. EC (European Commission) press releases are produced under the same EU-specific conditions as most EU documents are, i.e. multiple versions drafted and translated at the same time, non-native speakers drafting the documents etc. Culture has its own manner to construct and partition reality which is mirrored in its discourses, that is "modes of talking and thinking which can become ritualised" (Hatim and Mason, 1990). EU culture is no exception to that. In a corpus of aligned EC press releases an issue worth examining is whether the translation is affected by the different topics and discourses of the press releases.

2.3 Connectives: Relevance theory and Sentiment analysis

Connectives have been selected to be examined because they draw attention due to their status. According to Relevance Theory (Wilson and Sperber, 2002), the author produces his/her speech in such a way so that the reader will reach the speaker-intended interpretation with the least processing effort. The speaker, in order to achieve this, makes certain assumptions about the reader's background knowledge and, thus, expectations, and based on these assumptions formulates his/her discourse. From a relevance-theoretic perspective (Wilson and Sperber, 1993; Blakemore, 1987), connectives are not linking items, but devices whose meaning plays a part in the interpretation of an utterance. Among the different interpretations available, the hearer will decide which the speaker-intended one is, and connectives can facilitate the elimination of some of the available interpretations in order to achieve optimal relevance (Rouchota, 1998), i.e. the best possible interpretation for the hearer in terms of processing effort and effect.

Connectives have also been discussed in sentiment analysis. There is research which uses linguistic analysis and techniques to explore the sentiment of each sentence or phrase in a document. Meena and Prabhakar (2007) addressed the effects of conjunctions and sentence constructions in extracting sentiments associated with the phrases or sentences of reviews. Conjunctions are seen as crucial constituents when determining the polarity of a sentence. They found that, usually, either they alter the sentiment orientation to the opposite direction or they enhance the sentiment of the sentence.

Agarwal et al. (2008) involved in automatic sentiment analysis at sentence level in movie, car and book reviews observed that sentence structure has a fair contribution towards sentiment determination; conjunctions play a major role in defining the sentence structure. Their basic assumption is: "Not all phrases joined by a conjunct have same level of significance in overall sentiment determination".

3 Related tools

Parallel corpora are often used as linguistic resources in translation. Special tools have been designed to facilitate research in translation and multilingual parallel texts.

Callisto is a multilingual, multiplatform tool providing a set of “annotation services” (Day et al., 2004). Its standard components are textual annotation view and a configurable table display. Some of the tasks performed are automatic content extraction entity and relation detection, characterization and co-reference, temporal phrase normalization, named entity tagging, event and temporal expression tagging etc.

The IAMTC Project combines already existing facilities and newly developed ones and has developed an annotation tool for text manipulation. The Project involves the creation of multilingual parallel corpora with semantic annotation to be used in natural language applications (Farwell et al., 2008). Annotation includes dependency parsing, associating semantic concepts with lexical units, and assigning theta roles.

MULTEXT (Ide and Véronis, 1994) is a project involving the development of tools on the basis of ”software reusability”, and multilingual parallel corpora. It combines NLP and speech, and examines the possibilities for such a combination by harmonizing tools and methods from both areas. The annotation is performed with a segmenter, a morphological analyser, a part of speech disambiguator, an aligner, a prosody tagger, and post-editing tools. Thus, the annotated data provide information about syntax, morphology, prosody and the alignment of parallel texts.

Propbank is a project where a corpus is annotated with semantic roles for verb predicates (Choi et al., 2010). Annotation is performed with the help of Jubilee by simultaneously presenting syntactic and semantic information. The process is facilitated by Cornerstone, a user-friendly xml editor, customized to allow frame authors to create and edit frameset files.

Finally, there is ParaConc (Barlow, 2002) whose main characteristics are an alignment function, concordance search, search for specific words and their possible translations, corpus frequency and collocate frequency. But the tool has no annotating function.

These tools cannot fully meet the particularities of this research for the reasons discussed next.

4 Need for a new tool

The underlying factor that can bring the above different aspects and approaches together is an annotation tool that features certain specific characteristics that are hard to find all in one annotation tool. Coreference Annotator has those characteristics. In particular, a) uploading aligned texts already processed in an efficient alignment tool so as to achieve maximum alignment performance. The tool’s ability to have as input aligned documents allows a corpus builder to use a reliable external aligner of one’s own choice and then use the annotation scheme for the manual annotation of the aligned corpus; b) depicting the aligned texts in such an arrangement that each pair of aligned texts is clearly separated from the other pairs of aligned texts; each translation unit consisting of the source text segment and the target text segment in each pair of aligned texts is clearly and easily detectable from the other translation units. At the same time, it keeps its place in the text manifesting coherence and flow of text meaning in each language; c) allowing the location of possible translation equivalents in context of the instances of the linguistic items examined, always keeping the source text item and its target text equivalent in a close, binary relationship. This unfolds the variety of equivalents an item can have that may be either context dependent or context independent, and also highlights translation procedures and strategies; d) allowing the creation of a comparable profile at sentence level of the source text entry and the target text equivalent entry by entering accompanying information based on their context (distribution of the entries, collocations etc.) in the appropriate sections and fields of attributes – the source text entry and its equivalent text entry are seen comprehensively as a whole; e) displaying all the attribute sections and fields for each source text entry and its target text equivalent with one click to provide easy access which is important due to the large amount of data; f) allowing the examination of the target text in its own right to identify the cases, if any, of linguistic items under investigation that are present in the target text without being a translation equivalent of a source text entry; the annotation tool also provides for the creation of a profile for each target text addition entry; g) allowing the correlation of discourse topics with the frequency of the linguistic items and their translation equivalents in the two languages, and also with their microenvironments, thanks to the arrangement of the aligned texts; h) allowing the correlation of discourse topics, the frequency of the linguistic items and their translation equivalents, and the frequency of the items added in the target text; i) providing statistics based on the relationship of the source text entry and its target text equivalent where each result is fully and directly traceable in the corpus not only in terms of which pair of aligned texts it is found in but also in terms of its exact location in the pair, thus keeping track of text meaning and structure, and discourse; j) providing detailed statistics which allows the grouping of information of the profile of the entries for specialized analysis of results; k) producing tables of statistics exportable to widely commercial formats e.g. excel for further processing, e.g. SPSS. Such a sophisticated annotation tool allows multidisciplinary analysis. Finally, the tool has been implemented as a component of the Ellogon language engineering platform (Petasis et al., 2002), making extensive use of its infrastructure for the easy creation of annotation tools.

5 Corpus of Collections

This tool has been tested with a corpus of English-Greek press releases issued by the European Commission from 1/1/2007 to 1/1/2009. The corpus was drawn from the electronic text library of all EU press releases (RAPID)[1]. The criteria for text selection of that corpus are the availability of a Greek version and the currency of topics. The corpus consists of three thematic collections: the Environment, Agriculture, and Presidency Conclusions, which are further subdivided into thematic subcollections within each collection to make transparent the different discourses. The corpus has been aligned using the WinAlign alignment tool – an application of the SDL Trados 2007 suite. Exporting the aligned corpus in plain text format made it an appropriate input for the annotation tool which has been adapted to accommodate such input. The use of a long-standing professional alignment tool aims at achieving effective performance in the segmentation of the parallel texts at the level of equivalent sentences or text segments, i.e. translation units (SDL, 2007).

6 Annotation scheme

Annotation is conducted by associating attributes to the linguistic items. The annotation tool contains three sections of attribute fields. The first section is general and the most frequently used. In the first section, the focus is on the source text entry (ST EN) and the target text entry (TT EL) where the latter is considered the translation equivalent of the former in that context. The ST EN fields that follow relate to accompanying information of that token based on the particular context. The same goes for the TT EL fields. The next section, TT Addition, involves the addition of the items in question in the target texts. The third section, Context, involves the context of the texts. The original concept of that section is an attempt to map the differences emerging from the translation process between the two texts. There is great flexibility in designing the annotation scheme since using xml language allows the creation of different attributes and values or sections of attributes or the change of the existing attributes and values or sections of attributes.

6.1 Toolbar

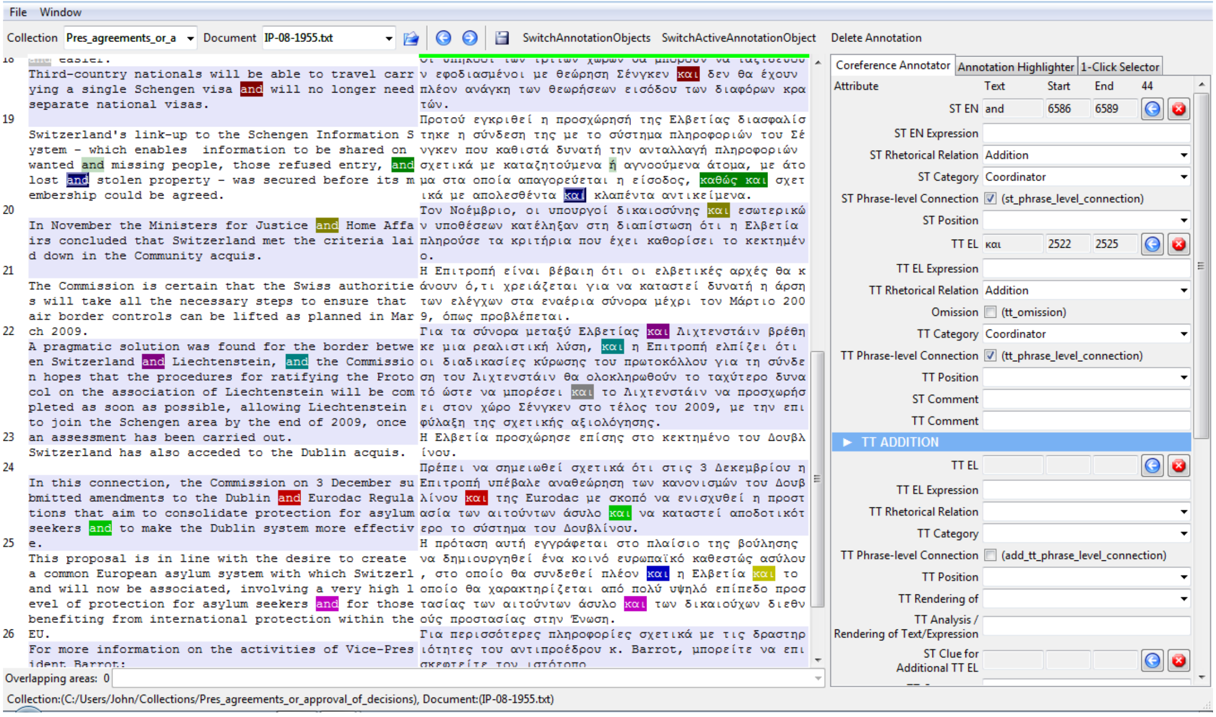
The toolbar is on the top of the screen (see Figure 1) where the collections and the filenames of the aligned documents of each collection are found. The arrow icons guide the annotator to the next or previous document of the collection. Few more icons facilitate managing the documents.

After selecting a collection and an aligned document, on the left side of the tool we can see the document in an aligned form – one column with the source text (ST) and one column with the target text (TT). The aligned document is presented in translation units, i.e. linked source and target text segments, with serial numbers for each unit for easier reference/retrieval when analysing a corpus. Also, to facilitate the visual separation of the translation units the background colour of the units alternates between white and light blue.

6.2 First section of attributes – General

On the right side of the tool, the three sections of attributes are presented. In the first section, the focus is on the source text entry (ST EN) and the target text entry (TT EL) where the latter is considered the translation equivalent of the former in that context. The ST EN and TT EL fields that follow relate to accompanying information of those tokens based on the particular context. When there is an arrow icon on the fields, there is a drop-down list of attributes to select. When an item is annotated the tool highlights it. Different annotated entries are highlighted with different colours but each ST EN entry has the same colour with its TT

[1] http://europa.eu/rapid/searchAction.do



Collection: Pres_agreements_or_a
Document: IP-08-1955.txt

18	easier. Third-country nationals will be able to travel carrying a single Schengen visa and will no longer need separate national visas.	Οι υπήκοοι των τρίτων χωρών θα μπορούν να ταξιδεύουν εφοδιασμένοι με θεώρηση Σένγκεν και δεν θα έχουν πλέον ανάγκη των θεωρήσεων εισόδου των διαφόρων κρατών.
19	Switzerland's link-up to the Schengen Information System – which enables information to be shared on wanted and missing people, those refused entry, and lost and stolen property – was secured before its membership could be agreed.	Προτού εγκριθεί η προσχώρησή της Ελβετίας διασφαλίστηκε η σύνδεση της με το σύστημα πληροφοριών του Σένγκεν που καθιστά δυνατή την ανταλλαγή πληροφοριών σχετικά με καταζητούμενα ή αγνοούμενα άτομα, με άτομα στα οποία απαγορεύεται η είσοδος, καθώς και σχετικά με απολεσθέντα και κλαπέντα αντικείμενα.
20	In November the Ministers for Justice and Home Affairs concluded that Switzerland met the criteria laid down in the Community acquis.	Τον Νοέμβριο, οι υπουργοί δικαιοσύνης και εσωτερικών υποθέσεων κατέληξαν στη διαπίστωση ότι η Ελβετία πληρούσε τα κριτήρια που έχει καθορίσει το κεκτημένο.
21	The Commission is certain that the Swiss authorities will take all the necessary steps to ensure that air border controls can be lifted as planned in March 2009.	Η Επιτροπή είναι βέβαιη ότι οι ελβετικές αρχές θα κάνουν ό,τι χρειάζεται για να καταστεί δυνατή η άρση των ελέγχων στα εναέρια σύνορα μέχρι τον Μάρτιο 2009, όπως προβλέπεται.
22	A pragmatic solution was found for the border between Switzerland and Liechtenstein, and the Commission hopes that the procedures for ratifying the Protocol on the association of Liechtenstein will be completed as soon as possible, allowing Liechtenstein to join the Schengen area by the end of 2009, once an assessment has been carried out.	Για τα σύνορα μεταξύ Ελβετίας και Λιχτενστάιν βρέθηκε μια ρεαλιστική λύση, και η Επιτροπή ελπίζει ότι οι διαδικασίες κύρωσης του πρωτοκόλλου για τη σύνδεση του Λιχτενστάιν θα ολοκληρωθούν το ταχύτερο δυνατό ώστε να μπορέσει και το Λιχτενστάιν να προσχωρήσει στον χώρο Σένγκεν στο τέλος του 2009, με την επιφύλαξη της σχετικής αξιολόγησης.
23	Switzerland has also acceded to the Dublin acquis.	Η Ελβετία προσχώρησε επίσης στο κεκτημένο του Δουβλίνου.
24	In this connection, the Commission on 3 December submitted amendments to the Dublin and Eurodac Regulations that aim to consolidate protection for asylum seekers and to make the Dublin system more effective.	Πρέπει να σημειωθεί σχετικά ότι στις 3 Δεκεμβρίου η Επιτροπή υπέβαλε αναθεώρηση των κανονισμών του Δουβλίνου και της Eurodac με σκοπό να ενισχυθεί η προστασία των αιτούντων άσυλο και να καταστεί αποδοτικότερο το σύστημα του Δουβλίνου.
25	This proposal is in line with the desire to create a common European asylum system with which Switzerland will now be associated, involving a very high level of protection for asylum seekers and for those benefiting from international protection within the EU.	Η πρόταση αυτή εγγράφεται στο πλαίσιο της βούλησης να δημιουργηθεί ένα κοινό ευρωπαϊκό καθεστώς ασύλου, στο οποίο θα συνδεθεί πλέον και η Ελβετία και το οποίο θα χαρακτηρίζεται από πολύ υψηλό επίπεδο προστασίας των αιτούντων άσυλο και των δικαιούχων διεθνούς προστασίας στην Ένωση.
26	For more information on the activities of Vice-President Barrot:	Για περισσότερες πληροφορίες σχετικά με τις δραστηριότητες του αντιπροέδρου κ. Barrot, μπορείτε να επισκεφτείτε τον ιστότοπο

Overlapping areas: 0

Collection:(C:/Users/John/Collections/Pres_agreements_or_approval_of_decisions), Document:(IP-08-1955.txt)

Attribute	Text	Start	End	44
ST EN	and	6586	6589	
ST EN Expression				
ST Rhetorical Relation	Addition			
ST Category	Coordinator			
ST Phrase-level Connection	☑ (st_phrase_level_connection)			
ST Position				
TT EL	και	2522	2525	
TT EL Expression				
TT Rhetorical Relation	Addition			
Omission	□ (tt_omission)			
TT Category	Coordinator			
TT Phrase-level Connection	☑ (tt_phrase_level_connection)			
TT Position				
ST Comment				
TT Comment				
▶ TT ADDITION				
TT EL				
TT EL Expression				
TT Rhetorical Relation				
TT Category				
TT Phrase-level Connection	□ (add_tt_phrase_level_connection)			
TT Position				
TT Rendering of				
TT Analysis / Rendering of Text/Expression				
ST Clue for Additional TT EL				

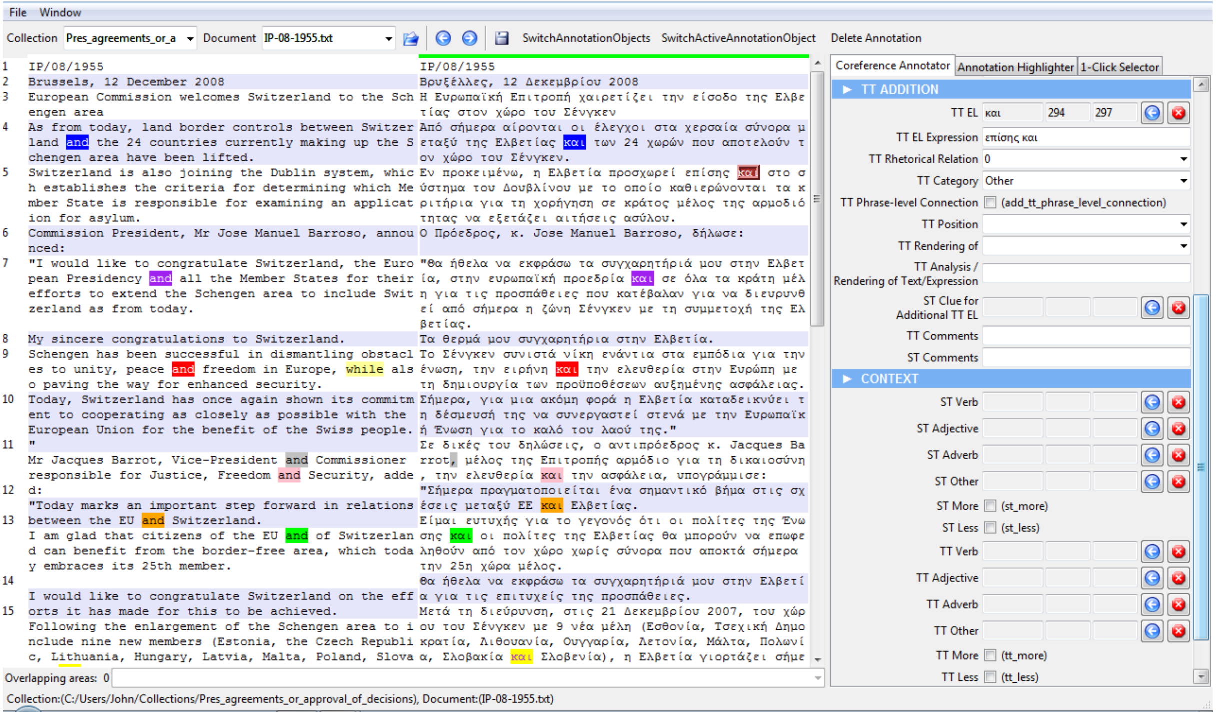
Figure 1: Toolbar and first section of Coreference Annotator – General.

EL equivalent entry. The fields ST EN/TT EL Expression accommodate cases where the ST EN/TT EL entries are part of an expression or form a collocation with the surrounding words. Each entry is also annotated for its rhetorical relation and category in that particular context. The values in these fields have been selected in relation to the connectives and discourse markers of interest. For cases where the discourse marker or connective has another function besides the linking one, the value "0" in the ST/TT Rhetorical Relation fields and the value "Other" in the ST/TT Category fields have been provided. There is also provision if a punctuation mark is in place of a TT EL entry. The checkbox of the ST/TT Phrase-level connection provides information about how often the ST and TT markers/connectives in question link predicates or non-predicates (noun phrases, adjectival phrases etc.) in their language respectively. Difference in the type of connection between the ST EN entry and its TT EL equivalent entry manifests different syntactic structures, and perhaps participant roles in the source and target languages. This in turn may reflect translation strategies e.g. shifts, transpositions, modulations etc. The ST/TT Position fields relate to the distribution of the tokens. When the ST EN entry and its TT EL equivalent are seen in parallel and a change in position is noted, then different thematic and rhematic structures, and focus may be reflected in the two languages. Omission of an ST EN entry in the target text is also checked. The last two fields, "ST Comment" and "TT Comment", allow comments by the annotator of the corpus that can be used either in revising or in analysing the corpus annotation.

An example can be a token of the additive conjunction 'and' (see Figure 1): This entry involves the token 'and', highlighted with blue colour in the translation unit 20. Based on its attributes, it is a conjunction of addition (ST Rhetorical Relation = "Addition"), a coordinator in particular (ST Category = "Coordinator"), and connects phrases (non-predicates) ("ST Phrase-level Connection" box checked). The token acting as its equivalent in the target text is και (kae) 'and', which is also a conjunction of addition (TT Rhetorical Relation = "Addition"), a coordinator (TT Category = "Coordinator"), and connects non-predicates ("TT Phrase-level Connection" box checked).

6.3 Second section of attributes – TT Addition

The next section, TT Addition, involves the addition of the items in question in the target texts (see Figure 2 – TT Addition). There are similar fields



Collection Pres_agreements_or_a Document IP-08-1955.txt

```
1   IP/08/1955                                             IP/08/1955
2   Brussels, 12 December 2008                             Βρυξέλλες, 12 Δεκεμβρίου 2008
3   European Commission welcomes Switzerland to the Sch    Η Ευρωπαϊκή Επιτροπή χαιρετίζει την είσοδο της Ελβε
    engen area                                             τίας στον χώρο του Σένγκεν
4   As from today, land border controls between Switzer    Από σήμερα αίρονται οι έλεγχοι στα χερσαία σύνορα μ
    land and the 24 countries currently making up the S    εταξύ της Ελβετίας και των 24 χωρών που αποτελούν τ
    chengen area have been lifted.                         ον χώρο του Σένγκεν.
5   Switzerland is also joining the Dublin system, whic    Εν προκειμένω, η Ελβετία προσχωρεί επίσης και στο σ
    h establishes the criteria for determining which Me    ύστημα του Δουβλίνου με το οποίο καθιερώνονται τα κ
    mber State is responsible for examining an applicat    ριτήρια για τη χορήγηση σε κράτος μέλος της αρμοδιό
    ion for asylum.                                        τητας να εξετάζει αιτήσεις ασύλου.
6   Commission President, Mr Jose Manuel Barroso, annou    Ο Πρόεδρος, κ. Jose Manuel Barroso, δήλωσε:
    nced:
7   "I would like to congratulate Switzerland, the Euro    "Θα ήθελα να εκφράσω τα συγχαρητήριά μου στην Ελβετ
    pean Presidency and all the Member States for their    ία, στην ευρωπαϊκή προεδρία και σε όλα τα κράτη μέλ
    efforts to extend the Schengen area to include Swit    η για τις προσπάθειες που κατέβαλαν για να διευρυνθ
    zerland as from today.                                 εί από σήμερα η ζώνη Σένγκεν με τη συμμετοχή της Ελ
                                                           βετίας.
8   My sincere congratulations to Switzerland.             Τα θερμά μου συγχαρητήρια στην Ελβετία.
9   Schengen has been successful in dismantling obstacl    Το Σένγκεν συνιστά νίκη ενάντια στα εμπόδια για την
    es to unity, peace and freedom in Europe, while als    ένωση, την ειρήνη και την ελευθερία στην Ευρώπη με
    o paving the way for enhanced security.                τη δημιουργία των προϋποθέσεων αυξημένης ασφάλειας.
10  Today, Switzerland has once again shown its commitm    Σήμερα, για μια ακόμη φορά η Ελβετία καταδεικνύει τ
    ent to cooperating as closely as possible with the     η δέσμευσή της να συνεργαστεί στενά με την Ευρωπαϊκ
    European Union for the benefit of the Swiss people.    ή Ένωση για το καλό του λαού της."
11  "                                                      Σε δικές του δηλώσεις, ο αντιπρόεδρος κ. Jacques Ba
    Mr Jacques Barrot, Vice-President and Commissioner     rrot, μέλος της Επιτροπής αρμόδιο για τη δικαιοσύνη
    responsible for Justice, Freedom and Security, adde    , την ελευθερία και την ασφάλεια, υπογράμμισε:
12  d:                                                     "Σήμερα πραγματοποιείται ένα σημαντικό βήμα στις σχ
    "Today marks an important step forward in relations    έσεις μεταξύ ΕΕ και Ελβετίας.
13  between the EU and Switzerland.                        Είμαι ευτυχής για το γεγονός ότι οι πολίτες της Ένω
    I am glad that citizens of the EU and of Switzerlan    σης και οι πολίτες της Ελβετίας θα μπορούν να επωφε
    d can benefit from the border-free area, which toda    ληθούν από τον χώρο χωρίς σύνορα που αποκτά σήμερα
    y embraces its 25th member.                            την 25η χώρα μέλος.
14                                                         Θα ήθελα να εκφράσω τα συγχαρητήριά μου στην Ελβετί
    I would like to congratulate Switzerland on the eff    α για τις επιτυχείς της προσπάθειες.
15  orts it has made for this to be achieved.              Μετά τη διεύρυνση, στις 21 Δεκεμβρίου 2007, του χώρ
    Following the enlargement of the Schengen area to i    ου του Σένγκεν με 9 νέα μέλη (Εσθονία, Τσεχική Δημο
    nclude nine new members (Estonia, the Czech Republi    κρατία, Λιθουανία, Ουγγαρία, Λετονία, Μάλτα, Πολωνί
    c, Lithuania, Hungary, Latvia, Malta, Poland, Slova    α, Σλοβακία και Σλοβενία), η Ελβετία γιορτάζει σήμε
```

Overlapping areas: 0

Collection:(C:/Users/John/Collections/Pres_agreements_or_approval_of_decisions), Document:(IP-08-1955.txt)

▶ TT ADDITION

TT EL και 294 297

TT EL Expression επίσης και

TT Rhetorical Relation 0

TT Category Other

TT Phrase-level Connection □ (add_tt_phrase_level_connection)

TT Position

TT Rendering of

TT Analysis / Rendering of Text/Expression

ST Clue for Additional TT EL

TT Comments

ST Comments

▶ CONTEXT

ST Verb

ST Adjective

ST Adverb

ST Other

ST More □ (st_more)

ST Less □ (st_less)

TT Verb

TT Adjective

TT Adverb

TT Other

TT More □ (tt_more)

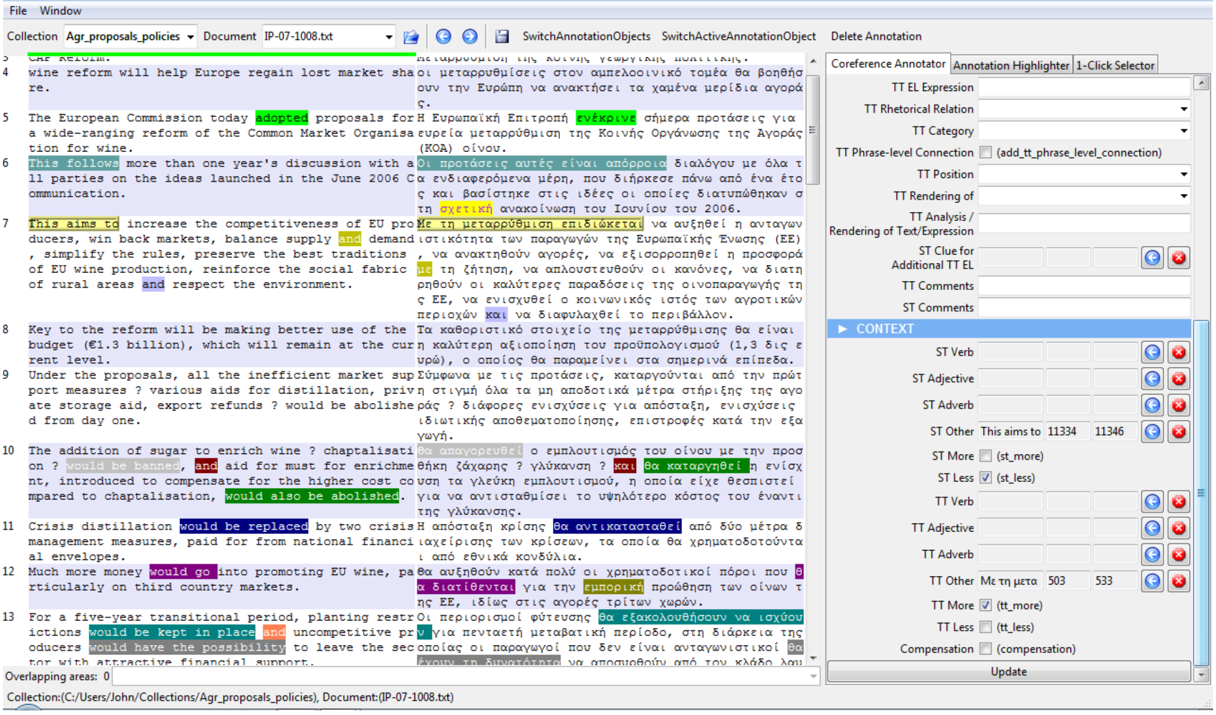
TT Less □ (tt_less)

Figure 2: TT Addition.

as in the first section of attributes. Because in this section of attributes the starting point is the target text, a couple of extra fields of attributes have been added: the "TT Rendering of" field which attempts to classify the category of the word/phrase in the ST, if any, that motivated the addition of the discourse marker/connective in the TT; the "TT Analysis/Rendering of Text/Expression" field where the ST word/phrase is entered. Finally, there is one more field, ST Clue for Additional TT EL. Practically, this and the previous field have a similar function. An example can be found in translation unit 5 (see Figure 2): According to the annotation, the TT EL entry και (kae) 'and' was added in translation unit 5, is not used as a conjunction (TT Rhetorical Relation=0) and performs a different function from coordination in the structure of the sentence (TT Category=Other).

6.4 Third section – Context

The third section involves the context of the texts (see Figure 3). The original concept of that section is an attempt to map the differences that emerge from the translation process. These differences can be grammatical e.g. a change in the tense of a verb form, semantic e.g. the choice of a slightly/a lot different semantically TT EL equivalent, pragmatic e.g. the choice of a completely different expression in the TT to render ST meaning, or lexical e.g. the addition or omission of a word/phrase in one of the two texts. The following pairs of fields have been designed: ST Verb (or verb phrase) – TT Verb (or verb phrase), ST Adjective (or adjectival phrase) – TT Adjective (or adjectival phrase), ST Adverb (or adverbial phrase) – TT Adverb (or adverbial phrase), ST Other – TT Other. The last pair involves differences that do not fall under any of the other pairs. Then the differences recorded can be evaluated compared with each other based on which of the two options – ST option or TT option – is more or less strong in meaning, more or less informative, more or less appellative, and more or less affective. Some of these differences between the two texts are mandatory driven by language restrictions, for instance, or optional driven by cultural preferences, register, politics etc. Either way, these differences create an effect to the reader. So under the ST fields there are two checkboxes ST More, ST Less and under the TT fields respectively TT More, TT Less. For each difference entered the relevant box is checked; ST entry evaluated as ST More or ST Less and TT equivalent evaluated as TT More or TT Less. There is one last checkbox in this section, Compensation, called after the translation strategy. Compensation refers to making up for the loss of meaning



Collection: Agr_proposals_policies
Document: IP-07-1008.txt

3	CAP Reform.	μεταρρύθμιση της κοινής γεωργικής πολιτικής.
4	wine reform will help Europe regain lost market share.	οι μεταρρυθμίσεις στον αμπελοοινικό τομέα θα βοηθήσουν την Ευρώπη να ανακτήσει τα χαμένα μερίδια αγοράς.
5	The European Commission today adopted proposals for a wide-ranging reform of the Common Market Organisation for wine.	Η Ευρωπαϊκή Επιτροπή ενέκρινε σήμερα προτάσεις για ευρεία μεταρρύθμιση της Κοινής Οργάνωσης της Αγοράς (ΚΟΑ) οίνου.
6	This follows more than one year's discussion with all parties on the ideas launched in the June 2006 Communication.	Οι προτάσεις αυτές είναι απόρροια διαλόγου με όλα τα ενδιαφερόμενα μέρη, που διήρκεσε πάνω από ένα έτος και βασίστηκε στις ιδέες οι οποίες διατυπώθηκαν στη σχετική ανακοίνωση του Ιουνίου του 2006.
7	This aims to increase the competitiveness of EU producers, win back markets, balance supply and demand, simplify the rules, preserve the best traditions of EU wine production, reinforce the social fabric of rural areas and respect the environment.	Με τη μεταρρύθμιση επιδιώκεται να αυξηθεί η ανταγωνιστικότητα των παραγωγών της Ευρωπαϊκής Ένωσης (ΕΕ), να ανακτηθούν αγορές, να εξισορροπηθεί η προσφορά με τη ζήτηση, να απλουστευθούν οι κανόνες, να διατηρηθούν οι καλύτερες παραδόσεις της οινοπαραγωγής της ΕΕ, να ενισχυθεί ο κοινωνικός ιστός των αγροτικών περιοχών και να διαφυλαχθεί το περιβάλλον.
8	Key to the reform will be making better use of the budget (€1.3 billion), which will remain at the current level.	Τα καθοριστικό στοιχείο της μεταρρύθμισης θα είναι η καλύτερη αξιοποίηση του προϋπολογισμού (1,3 δις ευρώ), ο οποίος θα παραμείνει στα σημερινά επίπεδα.
9	Under the proposals, all the inefficient market support measures ? various aids for distillation, private storage aid, export refunds ? would be abolished from day one.	Σύμφωνα με τις προτάσεις, καταργούνται από την πρώτη στιγμή όλα τα μη αποδοτικά μέτρα στήριξης της αγοράς ? διάφορες ενισχύσεις για απόσταξη, ενισχύσεις ιδιωτικής αποθεματοποίησης, επιστροφές κατά την εξαγωγή.
10	The addition of sugar to enrich wine ? chaptalisation ? would be banned, and aid for must for enrichment, introduced to compensate for the higher cost compared to chaptalisation, would also be abolished.	θα απαγορευθεί ο εμπλουτισμός του οίνου με την προσθήκη ζάχαρης ? γλύκανση ? και θα καταργηθεί η ενίσχυση τα γλεύκη εμπλουτισμού, η οποία είχε θεσπιστεί για να αντισταθμίσει το υψηλότερο κόστος του έναντι της γλύκανσης.
11	Crisis distillation would be replaced by two crisis management measures, paid for from national financial envelopes.	Η απόσταξη κρίσης θα αντικατασταθεί από δύο μέτρα διαχείρισης των κρίσεων, τα οποία θα χρηματοδοτούνται από εθνικά κονδύλια.
12	Much more money would go into promoting EU wine, particularly on third country markets.	Θα αυξηθούν κατά πολύ οι χρηματοδοτικοί πόροι που θα διατίθενται για την εμπορική προώθηση των οίνων της ΕΕ, ιδίως στις αγορές τρίτων χωρών.
13	For a five-year transitional period, planting restrictions would be kept in place and uncompetitive producers would have the possibility to leave the sector with attractive financial support.	Οι περιορισμοί φύτευσης θα εξακολουθήσουν να ισχύουν για πενταετή μεταβατική περίοδο, στη διάρκεια της οποίας οι παραγωγοί που δεν είναι ανταγωνιστικοί θα έχουν τη δυνατότητα να αποσυρθούν από τον κλάδο λαμ

Overlapping areas: 0

Collection:(C:/Users/John/Collections/Agr_proposals_policies), Document:(IP-07-1008.txt)

TT EL Expression
TT Rhetorical Relation
TT Category
TT Phrase-level Connection ☐ (add_tt_phrase_level_connection)
TT Position
TT Rendering of
TT Analysis / Rendering of Text/Expression
ST Clue for Additional TT EL
TT Comments
ST Comments
ST Verb
ST Adjective
ST Adverb
ST Other: This aims to | 11334 | 11346
ST More ☐ (st_more)
ST Less ☑ (st_less)
TT Verb
TT Adjective
TT Adverb
TT Other: Με τη μετα | 503 | 533
TT More ☑ (tt_more)
TT Less ☐ (tt_less)
Compensation ☐ (compensation)

Figure 3: Context.

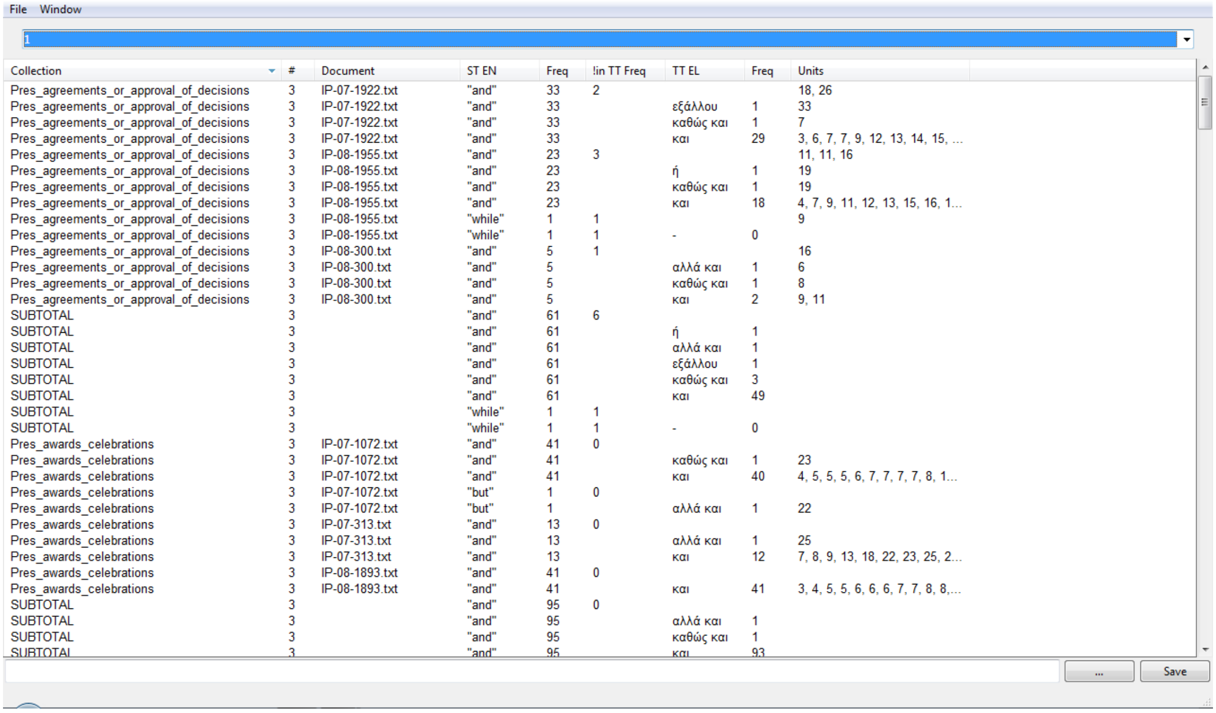
or effect in some part of the sentence in another part of that sentence or in a contiguous sentence (Newmark, 1988). This box is checked when the difference in context in the two texts is due to the translation strategy of compensation.

An example can be in translation unit 7 (see Figure 4): According to the annotation, the ST phrase ‘This aims to’ in translation unit 7, entered in the ST Other field is classified as ST Less compared to its TT equivalent phrase Με τη μεταρρύθμιση επιδιώκεται (Mae ti metarythmisi epidioketai) ‘With the reform it is aimed’. The reason is the act of referring in the English segment where the demonstrative pronoun ‘This’, a lexicalized deictic element or indexical, is clarified in the Greek segment with the nominal referent μεταρρύθμιση (metarythmisi) ‘reform’. So the TT phrase is more informative than the ST phrase. Because the foregrounded nominal in the TT phrase Με τη μεταρρύθμιση (Mae ti metarythmisi epidioketai) ‘With the reform it is aimed’ refers to the pronominal fronted in the ST, this is another factor which enhances the effect of the referring act in relation to the transposition between active and passive voice. Thus, the referring act prevails and classifies the TT phrase as TT More.

7 Statistics

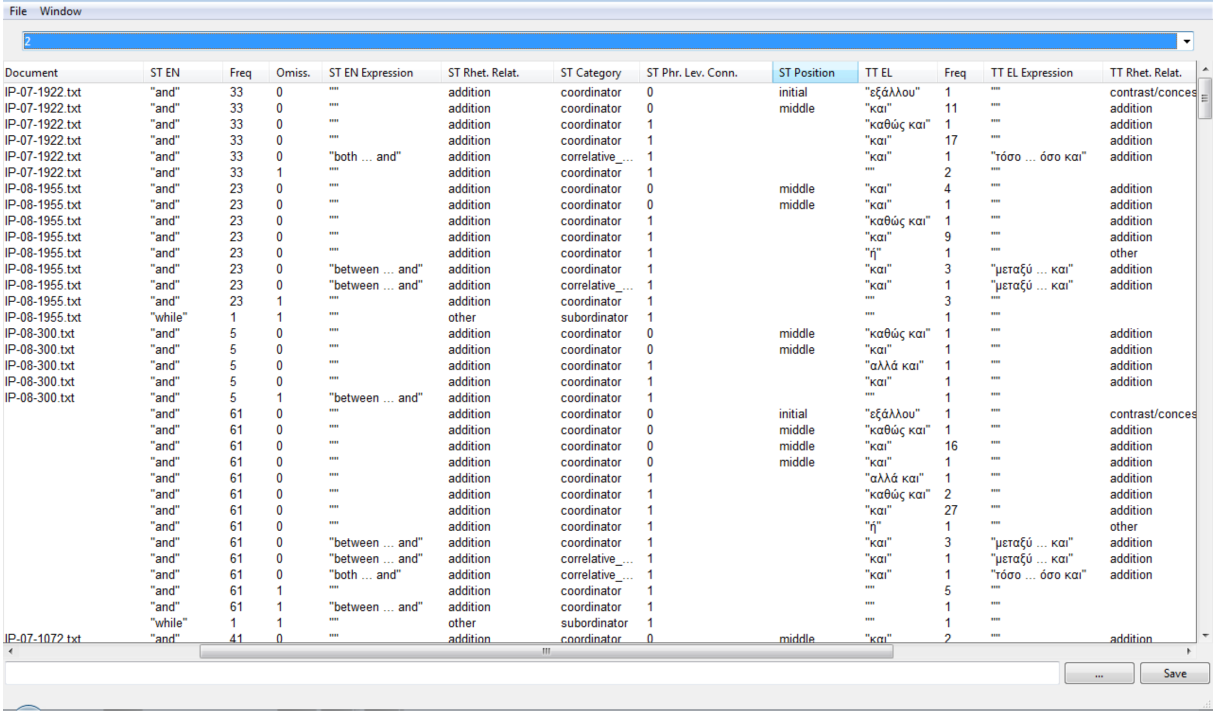
Detailed statistics tables are produced covering all possible search criteria. The findings are easily traceable in the corpus in terms of collection, aligned text and position of the translation unit where the item is found in the aligned text. In particular, three tables are generated. The first table (see Figure 4) presents all the source text tokens of interest per aligned document and collection, their frequency, their translation equivalents along with their own frequency, and cases of omission of the source text connectives/discourse markers in the target text. At the end of each collection, there is the subtotal of the frequency of source text connectives/discourse markers and their translation equivalents. After all the collections have been examined the table presents the total results of the total of collections. An important element is that next to each result there are the numbered translation units where the source text connective/discourse marker and its target text equivalent are found. This last feature allows easy retrieval of the translation unit, which ensures keeping track of text meaning and structure, and flow of discourse.

The second table (see Figure 5) presents grouped data based on the first section of at-



Collection	#	Document	ST EN	Freq	!in TT Freq	TT EL	Freq	Units
Pres_agreements_or_approval_of_decisions	3	IP-07-1922.txt	"and"	33	2			18, 26
Pres_agreements_or_approval_of_decisions	3	IP-07-1922.txt	"and"	33		εξάλλου	1	33
Pres_agreements_or_approval_of_decisions	3	IP-07-1922.txt	"and"	33		καθώς και	1	7
Pres_agreements_or_approval_of_decisions	3	IP-07-1922.txt	"and"	33		και	29	3, 6, 7, 7, 9, 12, 13, 14, 15, ...
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"and"	23	3			11, 11, 16
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"and"	23		ή	1	19
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"and"	23		καθώς και	1	19
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"and"	23		και	18	4, 7, 9, 11, 12, 13, 15, 16, 1...
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"while"	1	1			9
Pres_agreements_or_approval_of_decisions	3	IP-08-1955.txt	"while"	1	1	-	0	
Pres_agreements_or_approval_of_decisions	3	IP-08-300.txt	"and"	5	1			16
Pres_agreements_or_approval_of_decisions	3	IP-08-300.txt	"and"	5		αλλά και	1	6
Pres_agreements_or_approval_of_decisions	3	IP-08-300.txt	"and"	5		καθώς και	1	8
Pres_agreements_or_approval_of_decisions	3	IP-08-300.txt	"and"	5		και	2	9, 11
SUBTOTAL	3		"and"	61	6			
SUBTOTAL	3		"and"	61		ή	1	
SUBTOTAL	3		"and"	61		αλλά και	1	
SUBTOTAL	3		"and"	61		εξάλλου	1	
SUBTOTAL	3		"and"	61		καθώς και	3	
SUBTOTAL	3		"and"	61		και	49	
SUBTOTAL	3		"while"	1	1			
SUBTOTAL	3		"while"	1	1	-	0	
Pres_awards_celebrations	3	IP-07-1072.txt	"and"	41	0			
Pres_awards_celebrations	3	IP-07-1072.txt	"and"	41		καθώς και	1	23
Pres_awards_celebrations	3	IP-07-1072.txt	"and"	41		και	40	4, 5, 5, 5, 6, 7, 7, 7, 7, 8, 1...
Pres_awards_celebrations	3	IP-07-1072.txt	"but"	1	0			
Pres_awards_celebrations	3	IP-07-1072.txt	"but"	1		αλλά και	1	22
Pres_awards_celebrations	3	IP-07-313.txt	"and"	13	0			
Pres_awards_celebrations	3	IP-07-313.txt	"and"	13		αλλά και	1	25
Pres_awards_celebrations	3	IP-07-313.txt	"and"	13		και	12	7, 8, 9, 13, 18, 22, 23, 25, 2...
Pres_awards_celebrations	3	IP-08-1893.txt	"and"	41	0			
Pres_awards_celebrations	3	IP-08-1893.txt	"and"	41		και	41	3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 8,...
SUBTOTAL	3		"and"	95	0			
SUBTOTAL	3		"and"	95		αλλά και	1	
SUBTOTAL	3		"and"	95		καθώς και	1	
SUBTOTAL	3		"and"	95		και	93	

Figure 4: Statistics Table 1.



Document	ST EN	Freq	Omiss.	ST EN Expression	ST Rhet. Relat.	ST Category	ST Phr. Lev. Conn.	ST Position	TT EL	Freq	TT EL Expression	TT Rhet. Relat.
IP-07-1922.txt	"and"	33	0	""	addition	coordinator	0	initial	"εξάλλου"	1	""	contrast/conces
IP-07-1922.txt	"and"	33	0	""	addition	coordinator	0	middle	"και"	11	""	addition
IP-07-1922.txt	"and"	33	0	""	addition	coordinator	1		"καθώς και"	1	""	addition
IP-07-1922.txt	"and"	33	0	""	addition	coordinator	1		"και"	17	""	addition
IP-07-1922.txt	"and"	33	0	"both ... and"	addition	correlative_...	1		"και"	1	"τόσο ... όσο και"	addition
IP-07-1922.txt	"and"	33	1	""	addition	coordinator	1		""	2	""	
IP-08-1955.txt	"and"	23	0	""	addition	coordinator	0	middle	"και"	4	""	addition
IP-08-1955.txt	"and"	23	0	""	addition	coordinator	0	middle	"και"	1	""	addition
IP-08-1955.txt	"and"	23	0	""	addition	coordinator	1		"καθώς και"	1	""	addition
IP-08-1955.txt	"and"	23	0	""	addition	coordinator	1		"και"	9	""	addition
IP-08-1955.txt	"and"	23	0	""	addition	coordinator	1		"ή"	1	""	other
IP-08-1955.txt	"and"	23	0	"between ... and"	addition	coordinator	1		"και"	3	"μεταξύ ... και"	addition
IP-08-1955.txt	"and"	23	0	"between ... and"	addition	correlative_...	1		"και"	1	"μεταξύ ... και"	addition
IP-08-1955.txt	"and"	23	1	""	addition	coordinator	1		""	3	""	
IP-08-1955.txt	"while"	1	1	""	other	subordinator	1		""	1	""	
IP-08-300.txt	"and"	5	0	""	addition	coordinator	0	middle	"καθώς και"	1	""	addition
IP-08-300.txt	"and"	5	0	""	addition	coordinator	0	middle	"και"	1	""	addition
IP-08-300.txt	"and"	5	0	""	addition	coordinator	1		"αλλά και"	1	""	addition
IP-08-300.txt	"and"	5	0	""	addition	coordinator	1		"και"	1	""	addition
IP-08-300.txt	"and"	5	1	"between ... and"	addition	coordinator	1		""	1	""	
	"and"	61	0	""	addition	coordinator	0	initial	"εξάλλου"	1	""	contrast/conces
	"and"	61	0	""	addition	coordinator	0	middle	"καθώς και"	1	""	addition
	"and"	61	0	""	addition	coordinator	0	middle	"και"	16	""	addition
	"and"	61	0	""	addition	coordinator	0	middle	"και"	1	""	addition
	"and"	61	0	""	addition	coordinator	1		"αλλά και"	1	""	addition
	"and"	61	0	""	addition	coordinator	1		"καθώς και"	2	""	addition
	"and"	61	0	""	addition	coordinator	1		"και"	27	""	addition
	"and"	61	0	""	addition	coordinator	1		"ή"	1	""	other
	"and"	61	0	"between ... and"	addition	coordinator	1		"και"	3	"μεταξύ ... και"	addition
	"and"	61	0	"between ... and"	addition	correlative_...	1		"και"	1	"μεταξύ ... και"	addition
	"and"	61	0	"both ... and"	addition	correlative_...	1		"και"	1	"τόσο ... όσο και"	addition
	"and"	61	1	""	addition	coordinator	1		""	5	""	
	"and"	61	1	"between ... and"	addition	coordinator	1		""	1	""	
	"while"	1	1	""	other	subordinator	1		""	1	""	
IP-07-1072.txt	"and"	41	0	""	addition	coordinator	0	middle	"και"	2	""	addition

Figure 5: Statistics Table 2.

tributes. It includes the elements of the first statistics table enriched with the accompanying attributes of both source and target text entries. The results present linearly, focusing on the ST entry – TT equivalent entry pair, the attributes which accompany the pair. Every time an attribute of the pair changes, there is a different entry in the results. Again, information on the document, collection and translation unit where the pairs with the specific attributes are found satisfies any search criteria.

The third statistics table involves results from the second section of attributes – TT Addition. It follows the rationale of statistics table 2 (Figure 5) but it focuses only on the target text items that have been added without being a translation equivalent of the source text items in question. Statistics for the third section of attributes about Context has not been designed yet because this section of attributes has not been fully tested in the corpus.

8 Conclusion

The Coreference Annotator is an annotation tool which is user friendly in its operation. It gives the researcher the advantage of selecting an external alignment tool for aligning a corpus of parallel texts according to his/her needs. It allows great flexibility in the study of various linguistic items and the translation process at the same time providing, therefore, multiple levels of analysis. Thus the researcher works with a tool that is easily adjustable to his/her varied needs in relation with the annotation of bilingual data.

References

Ritesh Agarwal, T V Prabhakar, and Sugato Chakrabarty. 2008. “I Know What You Feel”: Analyzing the Role of Conjunctions in Automatic Sentiment Analysis. *GoTAL*, 5221(1):28–39.

Michael Barlow. 2002. ParaConc: concordance software for multilingual parallel corpora. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 20–24, Las Palmas, Canary Islands, Spain, May 29–31. European Language Resources Association.

Diana Blakemore. 1987. *Semantic constraints on relevance*. Blackwell.

Jinho D. Choi, Claire Bonial, and Martha Palmer. 2010. Multilingual propbank annotation tools: Cornerstone and jubilee. In *Proceedings of the NAACL HLT 2010 Demonstration Session*, HLT-DEMO ’10, pages 13–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Day, Chad McHenry, Robyn Kozierok, and Laurel Riek. 2004. Callisto : A configurable annotation workbench. In Maria Teresa Lino, Maria Francisca Xavier, Fatima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) 2004*, pages 2073–2076, Lisbon, Portugal, 5. ELRA, European Language Resources Association.

D. Farwell, S. Helmreich, B. Dorr, R. Green, F. Reeder, K. Miller, L. Levin, T. Mitamura, E. Hovy, O. Rambow, N. Habash, and A. Siddharthan, 2008. *Interlingual Annotation of Multilingual Text Corpora and FrameNet*. Moutin de Gruyter, Berlin.

Basil Hatim and Ian Mason. 1990. *Discourse and the translator*. Language in social life series. Longman.

Nancy Ide and Jean Véronis. 1994. Multext: Multilingual text tools and corpora. In *Proceedings of the 15th conference on Computational linguistics - Volume 1*, COLING ’94, pages 588–592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kaisa Koskinen. 2001. How to research EU translation? *Perspectives*, 9(4):293–300.

Kaisa Koskinen. 2004. Shared culture?: Reflections on recent trends in translation studies. *Target*, 16(1):143–156.

Vasilis Koutsivitis. 1994. *Theoria tis Metafrasis*. Ellinikes Panepistimiakes Ekdoseis, Athens, Greece.

Vasilis Koutsivitis. 2003. I proklisi tis polyglossias sti dievrimeni Evropaiki Enosi. In *Speech at the 4th Conference on Hellenic Language and Terminology, Athens*, Las Palmas, Canary Islands, Spain, October 30–31 and November 1st.

Arun Meena and T. V. Prabhakar. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *Proceedings of the 29th European conference on IR research*, ECIR’07, pages 573–580, Berlin, Heidelberg. Springer-Verlag.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice-Hall International, New York.

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Ion Androutsopoulos, and Constantine D. Spyropoulos. 2002. Ellogon: A New Text Engineering Platform. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 72–78, Las Palmas, Canary Islands, Spain, May 29–31. European Language Resources Association.

Villy Rouchota. 1998. Connectives, coherence and relevance. In Villy Rouchota and Andreas H. Jucker, editors, *Current Issues in Relevance Theory, Pragmatics & Beyond New Series*, volume 58, pages 11–58. John Benjamins Publishing Company.

SDL International, 2007. *WinAlign User Guide 2007*.

Anna Trosborg, editor. 1997. *Text Typology and Translation*. Benjamins Translation Library, 26. John Benjamins, Philadelphia.

Mara Tsoumari. 2008. The translation of EU texts and relevance. In E. Walaszewska, M. Kisielewska-Krysiuk, A. Korzeniowska, and M. Grzegorzewska, editors, *Relevant Worlds: Current Perspectives on Language, Translation and Relevance Theory*, pages 188–205. Cambridge Scholars Publishing, Newcastle, UK.

Deirdre Wilson and Dan Sperber. 1993. Linguistic form and relevance. *Lingua*, 90.

Deirdre Wilson and Dan Sperber. 2002. Relevance theory. *UCL Working Papers. Linguistics*, 14.