A New Annotation Tool for Aligned Bilingual Corpora

Georgios Petasis[1] and Mara Tsoumari[2]

[1] 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

[2] 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

Abstract. This paper presents a new annotation tool for aligned bilingual corpora, which allows the annotation of a wide range of information, ranging from information about words (such as part-of-speech tags or named-entities) to quite complex annotation schemas involving links between aligned segments, such as co-reference or translation equivalence between aligned segments in the two languages. The annotation tool is implemented as a component of the Ellogon language engineering platform, exploiting its extensive annotation engine, its cross-platform abilities and its linguistic processing components, if such a need arises. The new annotation tool is distributed with an open source license (LGPL), as part of the Ellogon language engineering platform.

Key words: Annotation tools, collaborative annotation, adaptable annotation schemas

1 Introduction

The huge amount of the available information on the Web has created the need of effective information extraction systems that are able to produce meta-data that satisfy user’s information needs. The development of such systems, in the majority of cases, depends on the availability of an appropriately annotated corpus in order to learn extraction models. The production of such corpora can be significantly facilitated by annotation tools. While a considerable number of annotation tools can be found in the literature [1,2], they are mostly targeting monolingual documents, lacking any support for aligned bilingual corpora. However, as parallel corpora are often used as linguistic resources in translation, some tools have been developed to facilitate research in translation and multilingual corpus analysis, including the following ones:

GATE [3] is a language engineering platform, which offers support for aligning corpora. Text alignment can be achieved at document, section, paragraph, sentence and word level. “Compound documents” are created by combining existing documents and by aligning various text segments between documents. It is unclear, however, whether the user can annotate information across the participating documents. Another popular tool is ParaConc [4] whose main characteristics are an alignment function, concordance search, search for specific words and their possible translations, corpus frequency and collocate frequency. However ParaConc offers no annotating facilities. A fairly recent tool is InterText[3], which is an editor for aligned parallel texts. It has been developed for the project InterCorp, in order to edit and manage alignments of multiple parallel language versions of texts at the level of sentences. However, similar to ParaConc and perhaps GATE, it does not support annotation of documents, only alignment between segments. Another parallel corpus alignment toolbox is Uplug [5], which is a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora. All these tools offer support for aligning segments in bilingual documents, but do not offer other annotation facilities, beyond alignment.

On the other hand, there exist tools that allow any type of linguistic annotation, but provide no special support for bilingual documents. Callisto is a multilingual, multi-platform tool providing a set of “annotation services” [6]. 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 [7]. Annotation includes dependency parsing, associating semantic concepts with lexical units, and assigning theta roles. MULTEXT [8] is a project involving the development of tools on the basis of “software re-usability”, 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. Finally, Propbank is a project where a corpus is annotated with semantic roles for verb predicates [9]. 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.

The tool presented in this paper is an attempt to narrow the gap between these two types of annotation tools. Allowing the alignment of bilingual, parallel

[3] http://wanthalf.saga.cz/intertext

documents at sentence level, the tool allows the annotation of any type of linguistic annotation on document pairs simultaneously, by presenting to the end user a synchronised view of both documents with aligned sentences next to each other.

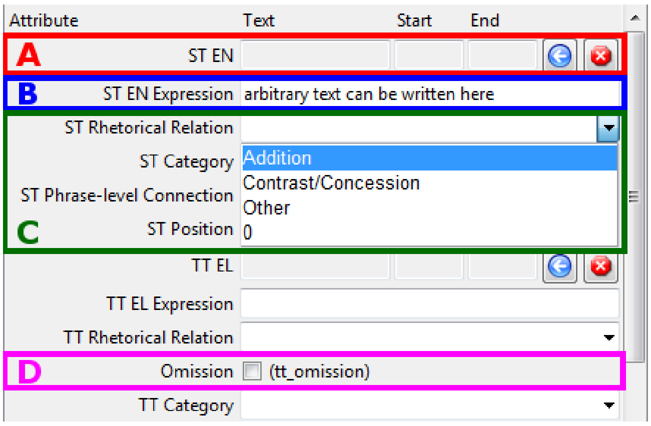
1.1 Motivation and features of the new annotation tool

The annotation tools presented in the previous section have been developed to cover specific and diverse needs, with each tool exhibiting different characteristics and capabilities, making difficult to find a single tool that concentrates the majority of features. Each of the tools presented in the previous section has significant features and capabilities but none of them is an all-in-one tool. Of course, the desired features of an annotation tool are closely related to the requirements of a specific annotation task, constituting the construction of a generic set of desired features quite difficult. 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. Being developed within the framework of wider research in the analysis of parallel texts from a translation point of view, the new annotation tool concentrates characteristics that cannot be found altogether in a single tool. In particular:
a) it imports aligned texts already processed in an efficient alignment tool, allowing a corpus builder to use an external aligner of one's own choice;
b) each pair (i.e. translation unit) of aligned texts is clearly separated from the other pairs. At the same time, they keep their place in the text manifesting coherence relations and flow of text meaning and discourse in each language;
c) the tool allows the location of possible translation equivalents within a specific context, 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 reveals translation procedures and strategies;
d) it allows the creation of a comparable profile at sentence level of the source text entry and its target text equivalent by entering accompanying information based on their context (distribution of the entries, collocations, etc.). The source text entry and its equivalent are seen comprehensively as a whole during the annotation process;
e) it displays all the attribute sections and fields for each source text entry and its target text equivalent, providing easy access with one click;
f) it allows the examination of the target text in its own right to identify cases, if any, of linguistic items that are present without being a translation equivalent of a source text entry;
g) it allows the correlation of discourse topics with the frequency of the linguistic items and their translation equivalents in the two languages, and also with their profile. Additionally, it allows both intra-linguistically and inter-linguistically analysis;
h) it provides detailed statistics [10], which allows the grouping of information

of the entries for specialized analysis of results; and the tables of statistics are exportable to widely commercial formats such as Microsoft Excel.

2 Reusing Ellogon's Annotation Engine

The Ellogon language engineering platform [11] offers an extensive annotation engine, allowing the construction of a wide range of annotation tools for both plain text and HTML documents. This annotation engine provides a wide range of features, including: *a*) cross-platform graphical user interface, *b*) use of standard formats, *c*) customized annotation schemata, *d*) automatic annotation, and *e*) comparison facilities to identify mismatches among independent annotations of the same document, or calculate inter-annotation agreement. Despite the fact that these features are not unique among the available annotation tools (i.e. most of these features are also supported by tools offered by Callisto, Wordfreak[4], GATE [3], MMAX2 [12], Knowtator [13], and AeroSWARM [14]), reusing an annotation engine allows for rapid and robust development of a new annotation tool, through the re-use of tested components. The annotation engine of the



ST EN

ST EN Expression: arbitrary text can be written here

ST Rhetorical Relation

ST Category: Addition / Contrast/Concession / Other

ST Phrase-level Connection

ST Position: 0

TT EL

TT EL Expression

TT Rhetorical Relation

Omission: □ (tt_omission)

TT Category

Fig. 1. Example of some annotation types, allowed in annotation schemas.

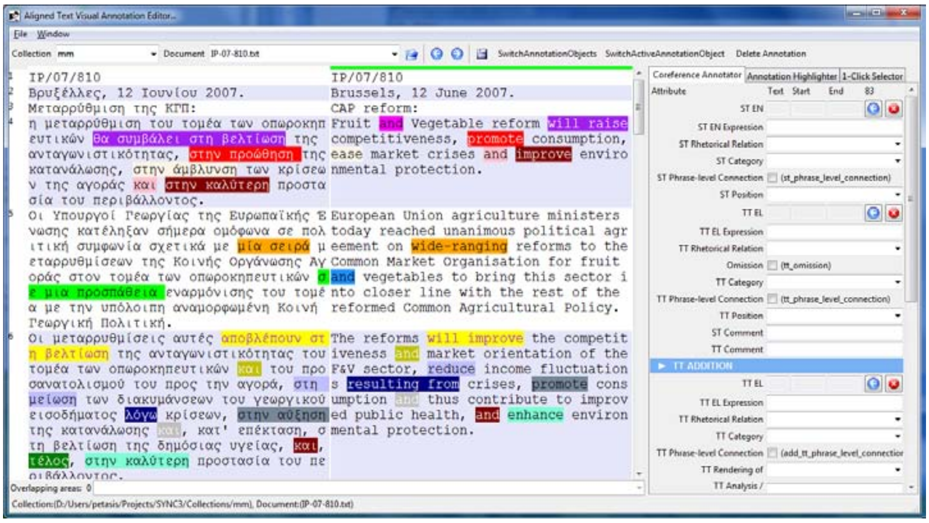
Ellogon language engineering platform is configurable through XML files that define annotation schemas. The tool reads the annotation schema from an XML file, and presents to the annotator a suitable GUI for annotating text segments. The XML annotation schema language provides a variety of types that can be annotated. While most available types, along with their visual representation in the GUI, can be found in [15] and [2], the types that relate to the grouping of several segments and other information in a single annotation to facilitate annotation of co-reference or other types of relations, are shown in the following

[4] http://wordfreak.sourceforge.net/

list: *a*) A **span** or **segment** (fig. 1-A), represented by a textual label (specified by the annotation schema), the text of the segment, its offsets, a button to fill in the segment from the current selection, and a button to clear the segment. *b*) A **description** (fig. 1-B), which the user can fill in with arbitrary text. Represented by a textual label and an entry widget, where arbitrary text can be entered. *c*) A **category** (fig. 1-C), selectable from a set of predefined categories by the annotation schema. Represented by a textual label and a combo-box widget, allowing the user to select a category from a set of predefined categories. *d*) A **boolean value** (fig. 1-D), denoting the presence or absence of an attribute. Represented by a textual label and a check-box widget.

2.1 The Aligned Bilingual Corpora Annotation Tool

After describing the annotation engine, the next step towards the creation of an annotation tool for aligned corpora is: *a*) to define a format for representing aligned bilingual corpora within the Ellogon platform, *b*) to extend the visualisation components to display correctly an aligned document, and *c*) to extend the annotation engine to operate on the extended visualisation components. A screenshot of the annotation tool can be seen in fig. 2. Aligned documents are displayed one next to the other, aligned at sentence level. The annotation schema used by the configuration shown in fig. 2 relates to the analysis of parallel texts from a translation point of view. Segments having the same colour between texts in the two languages in fig. 2 denote that they belong to the same annotation (group of features). Clicking on any of them enables the editing/modification of the relevant annotation, through the inputs on the right side of the tool.



Collection mm Document IP-07-810.txt

1 IP/07/810
2 Βρυξέλλες, 12 Ιουνίου 2007.
3 Μεταρρύθμιση της ΚΓΠ:
4 η μεταρρύθμιση του τομέα των οπωροκηπ
ευτικών θα συμβάλει στη βελτίωση της
ανταγωνιστικότητας, στην προώθηση της
κατανάλωσης, στην άμβλυνση των κρίσεω
ν της αγοράς και στην καλύτερη προστα
σία του περιβάλλοντος.
5 Οι Υπουργοί Γεωργίας της Ευρωπαϊκής Ε
νωσης κατέληξαν σήμερα ομόφωνα σε πολ
ιτική συμφωνία σχετικά με μία σειρά μ
εταρρυθμίσεων της Κοινής Οργάνωσης Αγ
οράς στον τομέα των οπωροκηπευτικών σ
ε μια προσπάθεια εναρμόνισης του τομέ
α με την υπόλοιπη αναμορφωμένη Κοινή
Γεωργική Πολιτική.
6 Οι μεταρρυθμίσεις αυτές αποβλέπουν στ
η βελτίωση της ανταγωνιστικότητας του
τομέα των οπωροκηπευτικών και του προ
σανατολισμού του προς την αγορά, στη
μείωση των διακυμάνσεων του γεωργικού
εισοδήματος λόγω κρίσεων, στην αύξηση
της κατανάλωσης και, κατ' επέκταση, σ
τη βελτίωση της δημόσιας υγείας, και,
τέλος, στην καλύτερη προστασία του πε
ριβάλλοντος.

IP/07/810
Brussels, 12 June 2007.
CAP reform:
Fruit and Vegetable reform will raise
competitiveness, promote consumption,
ease market crises and improve enviro
nmental protection.
European Union agriculture ministers
today reached unanimous political agr
eement on wide-ranging reforms to the
Common Market Organisation for fruit
and vegetables to bring this sector i
nto closer line with the rest of the
reformed Common Agricultural Policy.
The reforms will improve the competit
iveness and market orientation of the
F&V sector, reduce income fluctuation
s resulting from crises, promote cons
umption and thus contribute to improv
ed public health, and enhance environ
mental protection.

Overlapping areas: 0
Collection:(D:/Users/petasis/Projects/SYNC3/Collections/mm), Document:(IP-07-810.txt)

Fig. 2. The New Annotation Tool for Aligned Bilingual Corpora.

3 Usage Example: Connectives in Parallel Corpora

The main motivation for the development of this new annotation tool for aligned bilingual corpora has been the need to analyse the role of *connectives* in parallel corpora from both a linguistic and a translation perspective. To start with, the tool allows the researcher to build what we decided to call *parallel comparable corpora*. Parallel comparable corpora can be bilingual or multilingual collections (with more than one target language collections as translation of a source collection); be of the same genre or not; include source texts and their target texts; be grouped according to some textual resemblance – topics, text types etc. – although they cover different general topics; and be compared at various levels. For instance, a collection about the general topic agriculture is divided into subcollections with distinct text types and/or discourses and/or subtopics etc. The subcollections within the same collection can be contrasted with each other at the level of source text/language-intralinguistically or target text/language-intralinguistically. For instance if and how a linguistic form or a feature of context changes profile through different subcollections of the same collection. Also the subcollections can be contrasted at the level of translation-interlinguistically. For instance, if and how the translation of linguistic forms changes through the different subcollections. If more than one collection of a different general topic is included in the corpus, but with, at least to some extent, comparable subcollections, then similar subcollections of different collections can be compared both intralinguistically (compare only source texts among comparable subcollections or compare only target texts among comparable subcollections) and interlinguistically (compare how source texts are translated among comparable subcollections). If more than one genre is included then the comparison starts at the level of genres.

The tool is tested on a parallel comparable corpus, a special corpus of press releases of the European Commission, drawn from the electronic text library of all EU press releases (RAPID[5]), with two thematic categories-collections, Presidency with fifty pairs of English and Greek press releases, and Agriculture and rural development with fifty eight pairs of English and Greek press releases. Each of these thematic collections is further divided into separate thematic subcollections that can be comparable between the two collections, at least to some extent, despite the collections being of a different topic, i.e. Agriculture vs Presidency. Presidency collection is divided into six subcollections: Agreements or approval of decisions, Awards and celebrations, Visits and meetings, Proposals and policies, Various, Reports and surveys. Agriculture collection is divided into five subcollections: Agreements or approval of decisions, Proposals and policies, Reports and surveys, Approval of EU countries' plans, Warning and legal action. The names of the subcollections are comprehensive and cover a wide range of similar topics which have something in common, for instance somebody adopts something, the Commission proposes something, somebody is related to a meeting/conference/dialogue etc. The thematic subcollections in fact

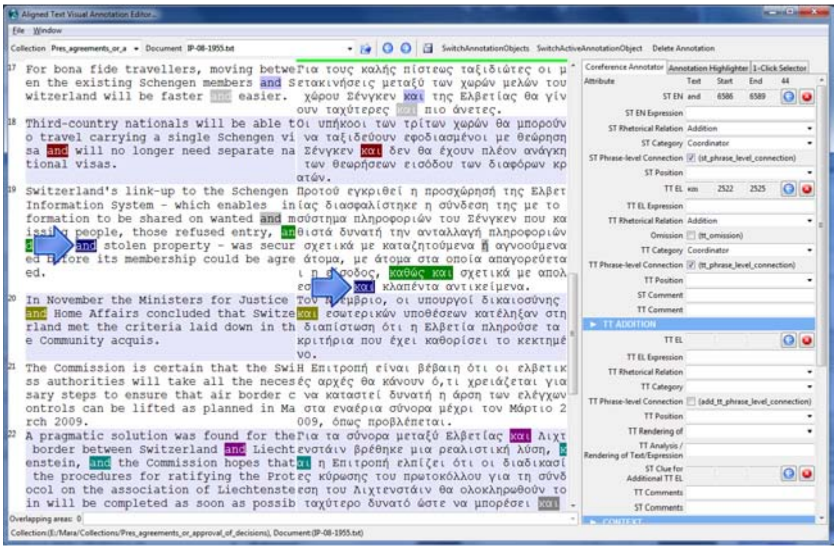
[5] http://europa.eu/rapid/searchAction.do

reflect different text types. Different text types may reflect different dominant discourses or functions or purposes. So far, the tool has processed both collections, Presidency and Agriculture by annotating the entries in question (adversative/contrastive/concessive discourse connectives and "and" connective). Analysis at this stage is focused on Presidency collection and its subcollections from a translation perspective. Findings show that the contrastive/concessive group of connectives keep their role in the Greek text with overwhelming persistence compared to "and". Only "yet" seems to differentiate itself from the rest of its group. Another finding is that the omission of discourse connectives in translation is not necessarily related to the large or small number of these connectives in the source texts. Having higher availability of a specific type of connective in a collection or subcollection is not necessarily related to higher omission rates of that type of connective. Findings from the addition of discourse connectives in the target texts show that contrast/concession is more persistent in the Greek press releases. Relating these findings with findings from a manual contextual analysis on a sample corpus of Agriculture collection the conclusion strengthens Sidiropoulou's [16,17] findings about the Greek reader viewing the world from a contrastive perspective. As to the question whether different text types affect the translation of these two groups of discourse connectives, One-Way ANOVA of "and" and its typical Greek translation equivalent "kai" – this pair was tested due to adequate amount of data – showed that there is a systematic influence of the text type on the frequency of the translation of "and" as 'και' . What determined the outcome of the One-Way Analysis of Variance is a particular subcollection Agreements or approval of decisions which has distinct features from the rest of the subcollections of that collection.

3.1 The Annotation schema

Annotation is conducted by associating attributes to the linguistic items. The devised annotation schema involves parallel documents in the English (EN) and Greek (EL) languages, and contains three sections of attribute fields: a) 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 of the TT EL entry. b) The next section, *target text addition (TT Addition)*, involves the addition of the items in question in the target texts, where there exists no connective or discourse marker in the source text. c) 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.

First Section of Attributes – General Section On the right side of the tool (fig. 3), the three sections of attributes defined by the annotation schema are presented. In the first section, the focus is on the *source text entry (ST EN)*



```
17 For bona fide travellers, moving betweΓια τους καλής πίστεως ταξιδιώτες οι μ
   en the existing Schengen members and Sετακινήσεις μεταξύ των χωρών μελών του
   witzerland will be faster and easier.  χώρου Σένγκεν και της Ελβετίας θα γίν
                                         ουν ταχύτερες και πιο άνετες.
18 Third-country nationals will be able tΟι υπήκοοι των τρίτων χωρών θα μπορούν
   o travel carrying a single Schengen vi να ταξιδεύουν εφοδιασμένοι με θεώρηση
   sa and will no longer need separate na Σένγκεν και δεν θα έχουν πλέον ανάγκη
   tional visas.                          των θεωρήσεων εισόδου των διαφόρων κρ
                                         ατών.
19 Switzerland's link-up to the Schengen Προτού εγκριθεί η προσχώρησή της Ελβετ
   Information System - which enables  inίας διασφαλίστηκε η σύνδεση της με το
   formation to be shared on wanted and mσύστημα πληροφοριών του Σένγκεν που κα
   issing people, those refused entry, anθιστά δυνατή την ανταλλαγή πληροφοριών
   d and stolen property - was secur σχετικά με καταζητούμενα ή αγνοούμενα
   ed before its membership could be agre άτομα, με άτομα στα οποία απαγορεύετα
   ed.                                    ι η είσοδος, καθώς και σχετικά με απολ
                                         ε[illegible] και κλαπέντα αντικείμενα.
20 In November the Ministers for Justice Τον Νοέμβριο, οι υπουργοί δικαιοσύνης
   and Home Affairs concluded that Switzeκαι εσωτερικών υποθέσεων κατέληξαν στη
   rland met the criteria laid down in th διαπίστωση ότι η Ελβετία πληρούσε τα
   e Community acquis.                    κριτήρια που έχει καθορίσει το κεκτημέ
                                         νο.
21 The Commission is certain that the SwiΗ Επιτροπή είναι βέβαιη ότι οι ελβετικ
   ss authorities will take all the necesές αρχές θα κάνουν ό,τι χρειάζεται για
   sary steps to ensure that air border c να καταστεί δυνατή η άρση των ελέγχων
   ontrols can be lifted as planned in Ma στα εναέρια σύνορα μέχρι τον Μάρτιο 2
   rch 2009.                              009, όπως προβλέπεται.
22 A pragmatic solution was found for theΓια τα σύνορα μεταξύ Ελβετίας και Λιχτ
    border between Switzerland and Liechtενστάιν βρέθηκε μια ρεαλιστική λύση, κ
   enstein, and the Commission hopes thatαι η Επιτροπή ελπίζει ότι οι διαδικασί
    the procedures for ratifying the Protες κύρωσης του πρωτοκόλλου για τη σύνδ
   ocol on the association of Liechtensteεση του Λιχτενστάιν θα ολοκληρωθούν το
   in will be completed as soon as possib ταχύτερο δυνατό ώστε να μπορέσει και
```

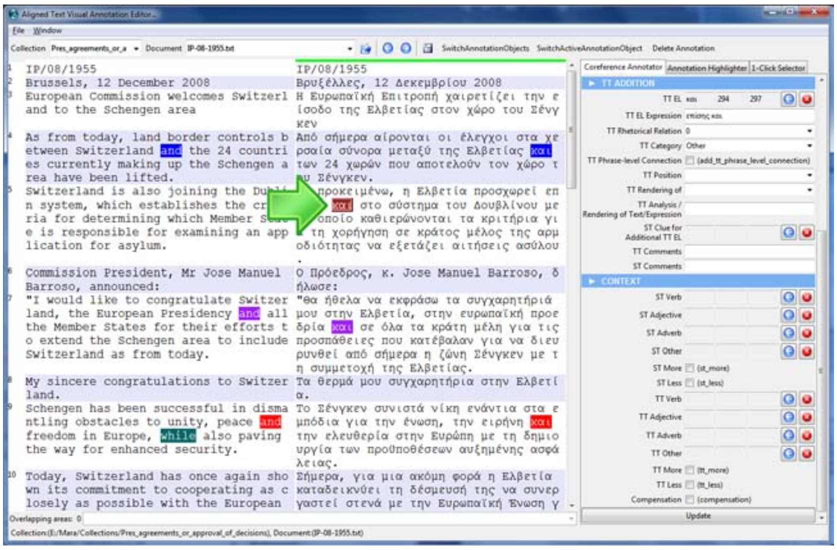
Attribute	Text	Start	End	44
ST EN	and	6586	6589	
ST EN Expression				
ST Rhetorical Relation	Addition			
ST Category	Coordinator			
ST Phrase-level Connection	☑ (st_phrase_level_connection)			
ST Position				
TT EL	και	2522	2525	
TT EL Expression				
TT Rhetorical Relation	Addition			
Omission	☐ (tt_omission)			
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				
TT Comments				
ST Comments				

Collection:(E:/Mara/Collections/Pres_agreements_or_approval_of_decisions), Document:(IP-08-1955.txt)

Fig. 3. First Section of Attributes – General Section.

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 annotated segments (tokens) based on the particular context. 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 except for 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. The check-box 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, i.e. 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 ("ST/TT Omission"). An example can be a token of the additive conjunction "and" (fig. 3): This entry involves the token "and", highlighted with blue colour in the translation unit 19. 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).



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 Switzerl and to the Schengen area	Η Ευρωπαϊκή Επιτροπή χαιρετίζει την ε ίσοδο της Ελβετίας στον χώρο του Σένγ κεν
4	As from today, land border controls b etween Switzerland and the 24 countri es currently making up the Schengen a rea have been lifted.	Από σήμερα αίρονται οι έλεγχοι στα χε ρσαία σύνορα μεταξύ της Ελβετίας και των 24 χωρών που αποτελούν τον χώρο τ ου Σένγκεν.
5	Switzerland is also joining the Du n system, which establishes the cr ria for determining which Member S e is responsible for examining an app lication for asylum.	προκειμένω, η Ελβετία προσχωρεί επ και στο σύστημα του Δουβλίνου με οποίο καθιερώνονται τα κριτήρια γι τη χορήγηση σε κράτος μέλος της αρμ οδιότητας να εξετάζει αιτήσεις ασύλου .
6	Commission President, Mr Jose Manuel Barroso, announced:	Ο Πρόεδρος, κ. Jose Manuel Barroso, δ ήλωσε:
7	"I would like to congratulate Switzer land, the European Presidency and all the Member States for their efforts t o extend the Schengen area to include Switzerland as from today.	"Θα ήθελα να εκφράσω τα συγχαρητήριά μου στην Ελβετία, στην ευρωπαϊκή προε δρία και σε όλα τα κράτη μέλη για τις προσπάθειες που κατέβαλαν για να διευ ρυνθεί από σήμερα η ζώνη Σένγκεν με τ η συμμετοχή της Ελβετίας.
8	My sincere congratulations to Switzer land.	Τα θερμά μου συγχαρητήρια στην Ελβετί α.
9	Schengen has been successful in disma ntling obstacles to unity, peace and freedom in Europe, while also paving the way for enhanced security.	Το Σένγκεν συνιστά νίκη ενάντια στα ε μπόδια για την ένωση, την ειρήνη και την ελευθερία στην Ευρώπη με τη δημιο υργία των προϋποθέσεων αυξημένης ασφά λειας.
10	Today, Switzerland has once again sho wn its commitment to cooperating as c losely as possible with the European	Σήμερα, για μια ακόμη φορά η Ελβετία καταδεικνύει τη δέσμευσή της να συνεργ αστεί στενά με την Ευρωπαϊκή Ένωση γ

Overlapping areas: 0

Collection:(E:/Mara/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)
Compensation □ (compensation)

Fig. 4. Second Section of Attributes – TT Addition.

Second Section of Attributes – TT Addition The next section, TT Addition, involves the addition of the items in question in the target texts (fig. 4). There are similar fields as in the first section of attributes. Since 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, the last two fields have a similar function providing distinct ways to enter information. An example can be found in translation unit 5 of fig. 4: according to the annotation, the TT EL entry και (kae) 'and' that 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").

Third Section of Attributes – Context The third section involves the context of the texts. The motivation of this section is an attempt to map the differences that emerge from the translation process. These differences can be: a) grammatical, i.e. a change in the tense of a verb form; b) semantic, i.e. the choice of a slightly/a lot different semantically TT EL equivalent; c) pragmatic, i.e. the choice of a completely different expression in the TT to render ST meaning; or d) lexical, i.e. the addition or omission of a word/phrase in one of the two texts. The following pairs of fields have been designed: *a*) "ST Verb" (or verb phrase) – "TT Verb" (or verb phrase), *b*) "ST Adjective" (or adjectival phrase) – "TT Adjective" (or adjectival phrase), *c*) "ST Adverb" (or adverbial phrase) – "TT Adverb" (or adverbial phrase), *d*) "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 check-boxes "ST More", "ST Less" and under the TT fields, "TT More" and "TT Less", respectively. 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". Finally, there is a check-box in this section, "Compensation", modelled after the translation strategy. Compensation refers to making up for the loss of meaning or effect in some part of the sentence, in another part of that sentence, or in a contiguous sentence [18]. This box is checked when the difference in context in the two texts is due to the translation strategy of compensation.

4 Conclusions and Future Work

In this paper we present a new annotation tool, which is able to annotate a wide range of information on aligned parallel corpora and parallel comparable corpora, implemented as a plug-in of the Ellogon language engineering platform, and distributed as open source. The annotation tool has been used in the context of analysing parallel/parallel comparable corpora from a translation point of view, concentrating mainly on the role of connectives. The annotation tool presented in this paper was proved extremely user friendly and robust in its operation, for the case studied. It offers to the researcher the advantage of selecting an external alignment tool for aligning a corpus of parallel texts according to his/her needs. In addition, it is very flexible when studying linguistic items and translation issues, and at the same time allows analysis pertaining to discourse, semiotics, ideology, culture etc. [10]. Thus the researcher works with a tool that is easily adjustable to his/her varied needs in relation with the annotation of bilingual data. As future work, we aim to integrate the aligned corpora annotation with the (semi) automatic annotation facilities offered by the Ellogon platform.

Acknowledgments The authors would like to acknowledge partial support from the FP7–ICT SYNC3 project.

References

1. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semant. **4** (January 2006) 14–28
2. Fragkou, P., Petasis, G., Theodorakos, A., Karkaletsis, V., Spyropoulos, C.: Boemie ontology-based text annotation tool. In: Proceedings of LREC 2008, ELRA (2008)
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M.A., Saggion, H., Petrak, J., Li, Y., Peters, W.: Text Processing with GATE. (2011)
4. Barlow, M.: ParaConc: concordance software for multilingual parallel corpora. In: Proceedings of LREC 2002. (2002)
5. Tiedemann, J.: ISA & ICA - two web interfaces for interactive alignment of bitexts. In: Proceedings of LREC 2006. (2006)
6. Day, D., McHenry, C., Kozierok, R., Riek, L.: Callisto : A configurable annotation workbench. In: Proceedings of LREC 2004. (2004)
7. Farwell, D., Helmreich, S., Dorr, B., Green, R., Reeder, F., Miller, K., Levin, L., Mitamura, T., Hovy, E., Rambow, O., Habash, N., Siddharthan, A. In: Interlingual Annotation of Multilingual Text Corpora and FrameNet, Berlin (2008)
8. Ide, N., Véronis, J.: Multext: Multilingual text tools and corpora. In: Proceedings of COLING 1994 - Volume 1, ACL (1994) 588–592
9. Choi, J.D., Bonial, C., Palmer, M.: Multilingual propbank annotation tools: Cornerstone and jubilee. In: Proceedings of the NAACL HLT 2010 Demo Session, ACL (2010) 13–16
10. Tsoumari, M., Petasis, G.: Coreference Annotator - A new annotation tool for aligned bilingual corpora. In: Proceedings of AEPC 2, RANLP 2011. (2011)
11. Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos, I., Spyropoulos, C.D.: Ellogon: A New Text Engineering Platform. In: Proceedings of LREC 2002. (2002)
12. Müller, C., Strube, M.: Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., Mukherjee, J., eds.: Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods. Peter Lang, Frankfurt a.M., Germany (2006) 197–214
13. Ogren, P.V.: Knowtator: A protégé plug-in for annotated corpus construction. In Moore, R.C., Bilmes, J.A., Chu-Carroll, J., Sanderson, M., eds.: HLT-NAACL, The Association for Computational Linguistics (2006)
14. Corcho, O.: Ontology based document annotation: trends and open research problems. Int. J. Metadata Semant. Ontologies **1** (January 2006) 47–57
15. Petasis, G.: The SYNC3 Collaborative Annotation Tool. In: Proceedings of LREC 2012. (2012)
16. Sidiropoulou, M.: Contrast in english and greek newspaper reporting: A translation perspective. In: Proceedings of 8^{th} International Symposium on English & Greek: Description and/or Comparison of the two languages, Thessaloniki, Greece, School of English, Aristotle University (1994)
17. Sidiropoulou, M.: Linguistic Identities through Translation. Volume 23 of Approaches to Translation Studies. Rodopi B.V., Amsterdam/New York (2004)
18. Newmark, P.: A Textbook of Translation. Prentice-Hall International, New York (1988)