Deliverable 7.1

Report on survey results and evaluation plan

The project has received funding from the European Community (ICT-PSP 4th call)
under Grant Agreement n° 270915.

DELIVERABLE

Project Acronym:	**Bologna**
Grant Agreement number:	**270915**
Project Title:	**Bologna Translation Service**

D7.1 - Report on survey results and evaluation plan

Revision: 4

Authors:

Joeri Van de Walle (Cross Language)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	
C	Confidential, only for members of the consortium and the Commission Services	X

The project has received funding from the European Community (ICT-PSP 4th call)
under Grant Agreement n° 270915.

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
1	16/8/2011	Joeri Van de Walle	Cross Language	First draft
2	19/8/2011	Heidi Depraetere	Cross Language	Edits
3	15/9/2011	Andy Way	ALS	Edits
4	19/9/2011	Heidi Depraetere	Cross Language	Final edits

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Introduction

This deliverable consists of two parts:

- the survey results, and
- the evaluation plan

In the section on survey results we will first discuss the aim and set-up of the survey. Next we will list and discuss the most striking findings. All questions and responses to the survey can be found in the appendix to this deliverable.

In the section on the evaluation plan we will describe our approach to evaluation.

Survey Results

Aim and Set-up

To collect feedback on user requirements for the Bologna Translation Service a survey was created and put up on the Bologna website (http://www.bologna-translation.eu/survey). All user group members, but also all occasional website visitors, were invited to take the survey.

The main goal of the survey was to hear from users about the functional requirements of the service, including: What do potential users think are required features for the translation platform? Which file formats would they like to be able to translate? Will they provide post-editors of their own or would they expect the Bologna Translation Service to have a pool of post-editors that they could use? For the complete list of questions asked in the survey, please refer to "Appendix A - Survey Questions".

The number of participants has not been that large so far but this was to be expected as the project had only just started. To date, we have had 12 responses, mostly from user group members. Because of the somewhat limited response, we should be careful about drawing conclusions from the survey. With the additional effort that will be put into expanding the user group, we hope to obtain additional feedback.

With this in mind, the results discussed below should be seen as an intermediary report rather than as a final one. We hope to be able to collect more responses as the project continues and anticipate having our initial findings confirmed by feedback from additional informants.

In the next section, we will discuss the preliminary findings of the survey.

Survey Highlights

For a number of the questions asked, users' preferences were not so outspoken. This may be due to the limited number of responses obtained so far.

However, for quite a few questions, users seemed to have some clear preferences. For instance, most users seem to agree on the following aspects of the service:

- Registration procedure,
- Language requirements,
- Post-editing requirements,
- Request size,
- Interaction with the service, and
- File format requirements

We discuss each of these in turn below.

Registration Procedure

It seems that most informants tend to agree that users should be able to register themselves online without help from the Bologna administrator (question 1). In addition, it is clear that companies/ universities want to keep control over which users they allow to make translation requests. A majority of 67% of the respondents are of the opinion that new registrations should be approved by the company/university administrator before users can start making translation requests (question 2).

Language Requirements

All respondents require translation from their native language into English (100%). The need to translate from multiple languages or to translate into languages other than English is currently completely absent, although a larger community of users can be expected to demonstrate some such requirement.

Post-editing Requirements

It is clear that the majority of respondents want to deal with post-editing internally. Either requesters of translations will post-edit themselves (75%), or they will turn to colleagues to do it for them (58%).

One third of the respondents would like to see a pool of post-editors as part of the service offered by the Bologna Translation Service.

A majority of 58% of the respondents feel that selecting post-editors for a job should be the responsibility of the company or university administrator. Respondents see ‘speciality’ as the main criterion for deciding which post-editor is most fit for the job (75%).

Another interesting finding is that 73% of the respondents see no need for the translation to be reviewed. After post-editing has been performed they would like to obtain the translation immediately.

Request Size

A clear majority of respondents (64%) expect the size of the translation requests to be small enough for one post-editor to handle.

Interaction with the Service

When presented with the options of sending e-mail or submitting request via a web interface, a large majority of respondents (67%) have a preference for the latter.

On the other hand, 83% of the respondents would like to receive a notification by e-mail when a request has been processed. They then intend to go online to download the requests. At the same time, half of the respondents could live with other delivery options: either no notification being sent, or receiving the translation by e-mail once it has been completed.

File Format Requirements

It seems like text-based formats are the most important requirement for the respondents. Everyone wants to be able to translate plain text, but also translation of HTML (92%), URL (83%), and XML (83%) are high on the priority list.

The requirement for translating Microsoft Office formats or PDF seems to be a lot smaller.

Evaluation Plan

In this section we will describe our approach to evaluation. The progress and results of the Bologna project will be evaluated from different perspectives:

- usability and utility to users,
- translation quality, and
- cost savings potential.

First and foremost, the service will be evaluated from a *user* perspective. The project will deploy methods that will try to assess the usability and utility of the service. The survey discussed in the previous section of this document plays an important role in this, but in addition we plan to perform more structured evaluations to evaluate this part of the service.

For the structured usability evaluation 30 study programmes per language will be selected and translated. Members of the user group will be asked a number of questions about the translations in order to inform us as to how useful the translations are in the context that the users intend to use them. The objective here is to think about real-life scenarios in which users of the service may rely on machine translation (MT), and to see to what degree MT succeeds in assisting the user in the way that he expects. Questions and translations will be presented to users as an online survey, allowing for easy processing of the results.

This structured usability evaluation will happen at two stages in the project: once at M12, with the output produced by the first versions of the engines, and again at the end of the project with output of the Bologna fine-tuned advanced systems.

It is hard to set a goal for the usability results, but indirectly they will be measured by looking at the number of users actively using the service together with the number of requests processed. Of course, these figures will only become apparent after the system has been put into production.

In addition, the MT systems will be evaluated from a *translation quality* perspective. Another objective of the evaluation plan is to provide objective information regarding the quality of the translations that the systems produce. Whereas the usability evaluation is context-based and focuses on complete translations of course programmes, translation quality evaluation will evaluate the quality of the translation by inspecting the translations of individually translated segments.

Translation quality will be measured in different ways. Firstly, it will be measured by using automatic evaluation scripts. Measures such as BLEU, NIST and TER will be applied by system developers to fixed development and test sets on a continual basis throughout the lifetime of the project in order to track improvements in translation quality of the systems being built. The basis for this automatic evaluation will be a test set of 1000 randomly selected segments per language. The goal the project sets itself is to achieve a minimum reduction in TER of 10% on these 300 segments when comparing the translation quality that the engines produce at the mid-project point against the quality produced at the end of the project.

Secondly, translation quality will be measured by having human evaluators judge the quality of the translations produced. The basis for this human evaluation will be an evaluation set of 300 randomly selected segments per language. The set used for the human evaluation will be a subset extracted from the set used for automatic evaluation.

The human evaluation will consist of three parts:

- adequacy and fluency evaluation,
- a benchmarking evaluation, and
- an error classification

The adequacy measurement is based on evaluation metrics as developed by the Defense Advanced Research Projects Agency (DARPA). Evaluators will score each segment in the evaluation set on a five to one degrading scale where five indicates that all of the meaning is present in the MT output and one indicates that little or no meaning is present in the MT output. Using a similar degrading scale for fluency, five would indicate 'completely fluent' with one indicating 'word salad'.

The second part of the human evaluation is a benchmarking exercise. Evaluators will be presented with an evaluation set for ranking the MT output of each segment from the integrated system and MT output from two other publicly available MT systems including rule-based MT and statistical MT.

The third part of the human evaluation is based on error categorisation and consists of identifying and classifying MT errors. Each segment in the evaluation set will be checked for errors and each error will be assigned a category. A distinction is made between errors caused by mistakes in the source text and those caused by the translation engine. Error categorisation of this type will permit the MT systems to be improved, either in the system-building phase, or in an automatic post-editing phase using regular expressions.

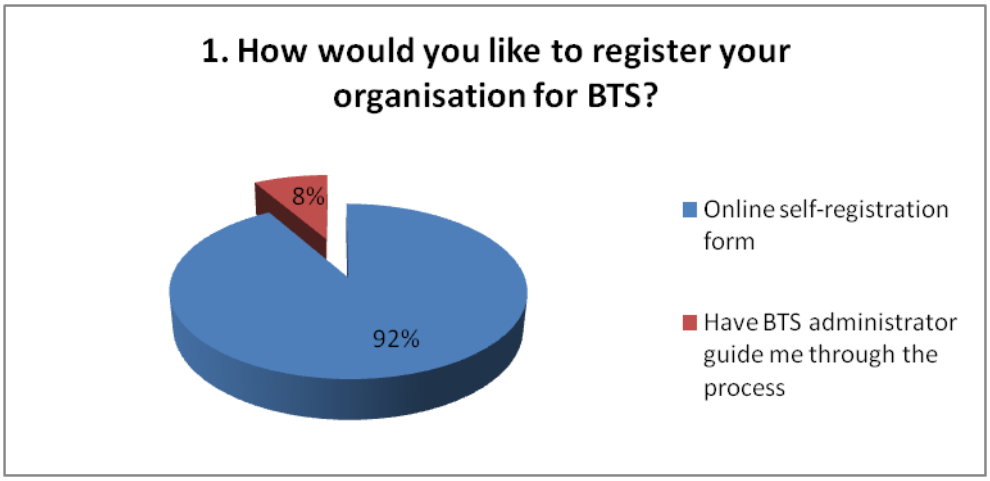
Part one of the human evaluation will be performed twice in the course of the project: once when the first version of the advanced systems has been released (between M10 and M15) and once at the end of the project. Parts two and three of the human evaluation will only be performed at the end of the project.

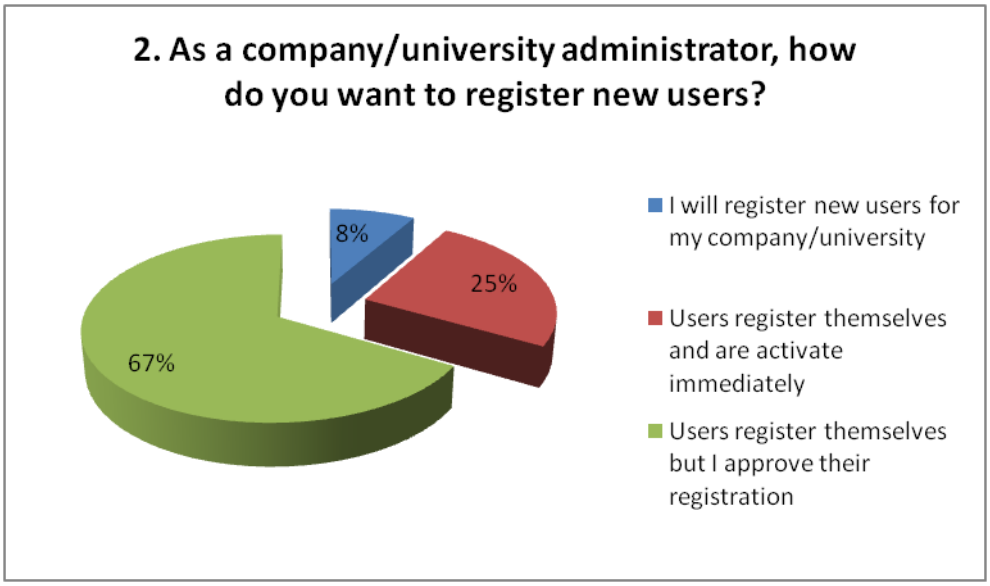
Third, the translation systems will be evaluated from an *economic* perspective. We will assess to what extent productivity increases (with concomitant cost savings) may be obtained by using MT as an aid to increase the speed of human translation.

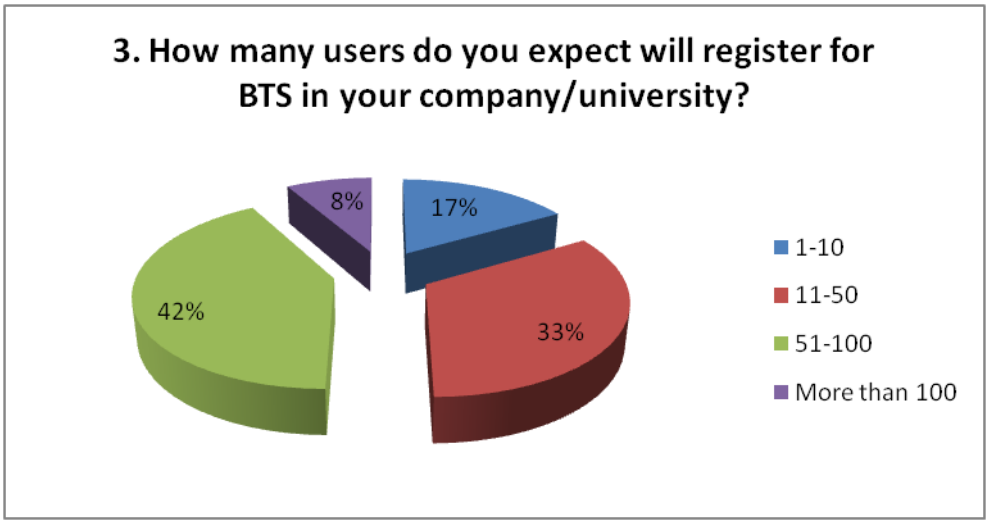
This productivity evaluation involves post-editing the MT output. The same set of 300 segments that is used for the human quality evaluation will be used and will be extended with an additional 200 segments to be translated from scratch. The time that the linguist spends on post-editing, reviewing or translating each of the segment types will be measured in order to calculate and compare the average throughput for each of the categories. The results will highlight potential productivity increase and cost savings. We anticipate using multiple post-editors and computing inter-rater agreement.

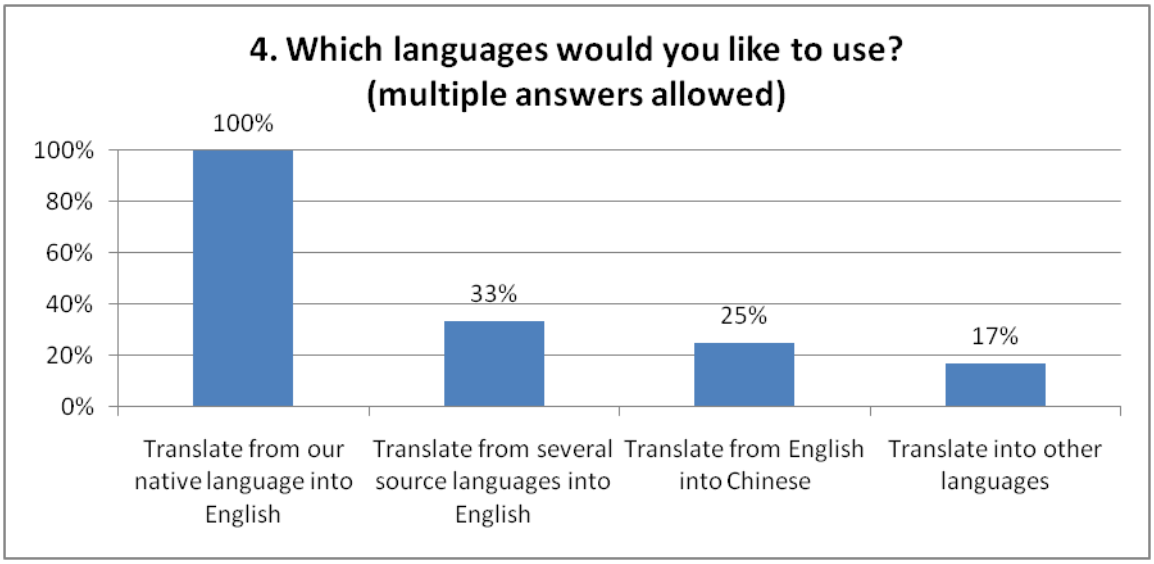
The productivity evaluation will be carried out once at the end of the project, and the objective is to reach a minimum of 10% productivity increase when looking at the throughput obtained by post-editing MT output compared to that obtained by translating from scratch.

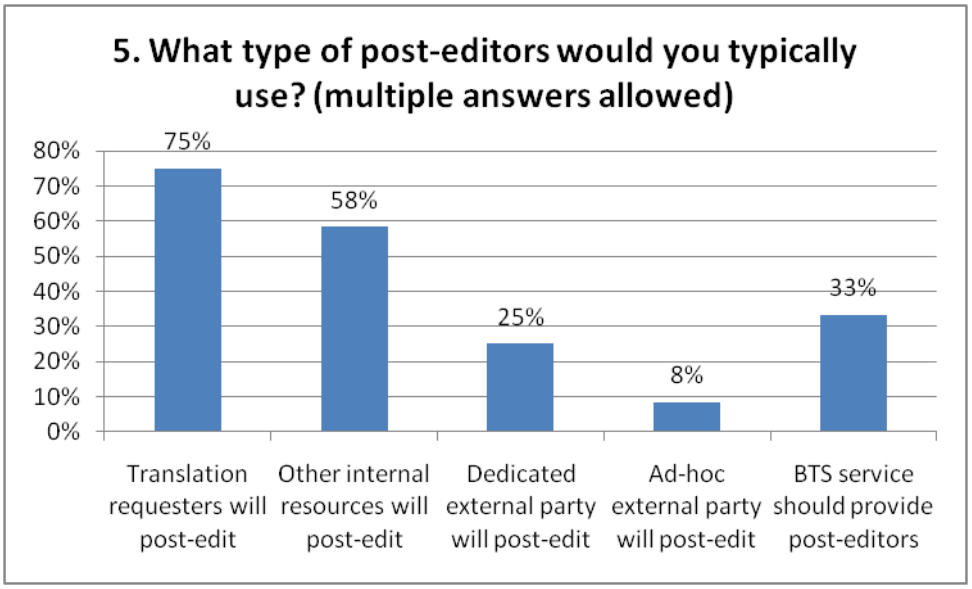
Appendix A - Survey Questions

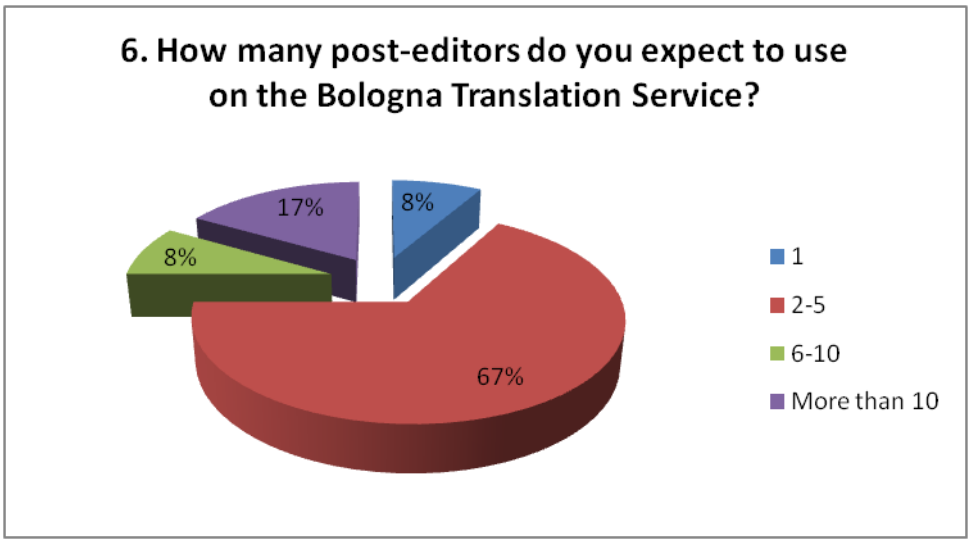


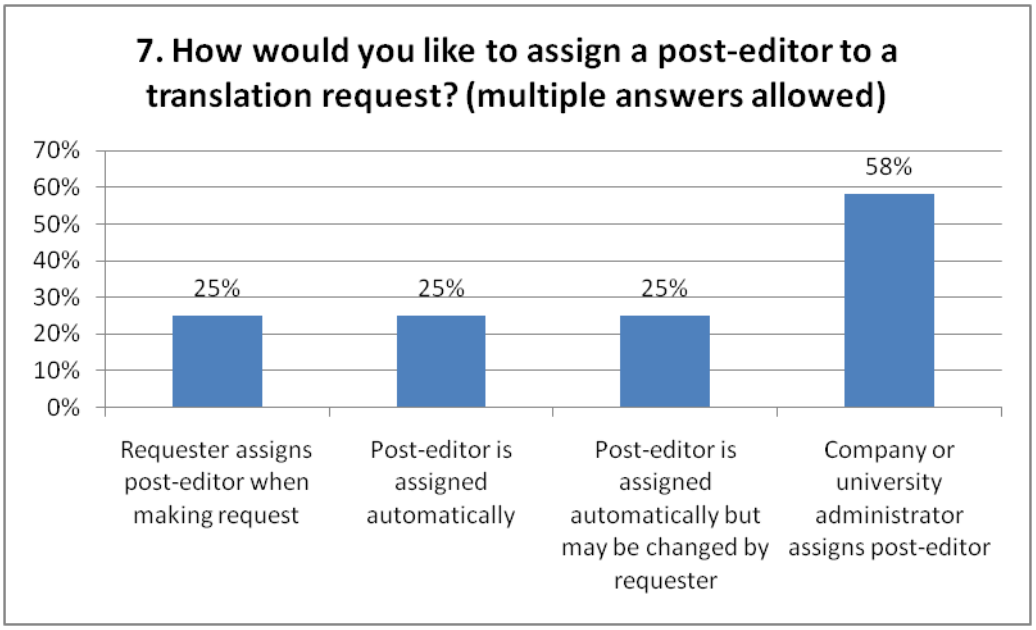


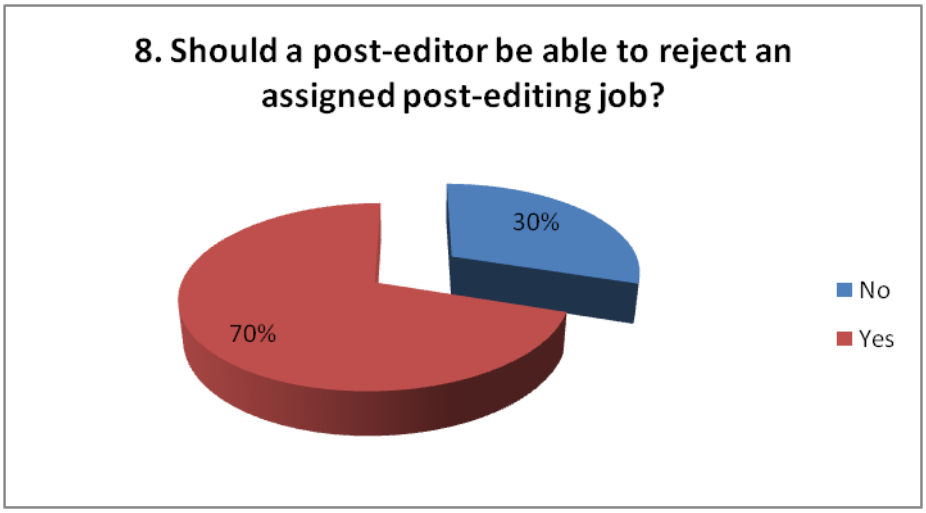


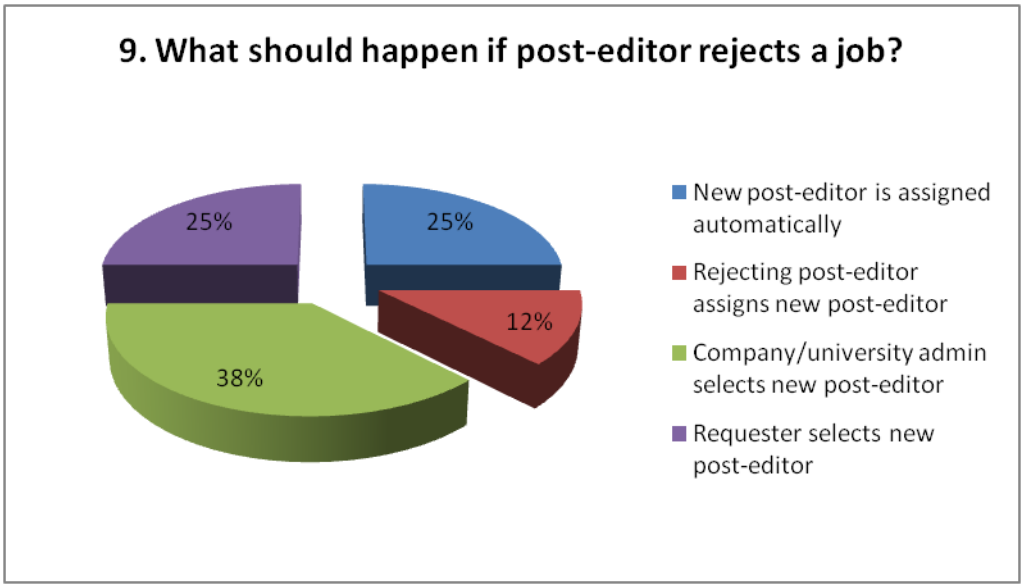

4. Which languages would you like to use? (multiple answers allowed)
100%
80%
60%
40%
20%
0%
100%
33%
25%
17%
Translate from our native language into English
Translate from several source languages into English
Translate from English into Chinese
Translate into other languages

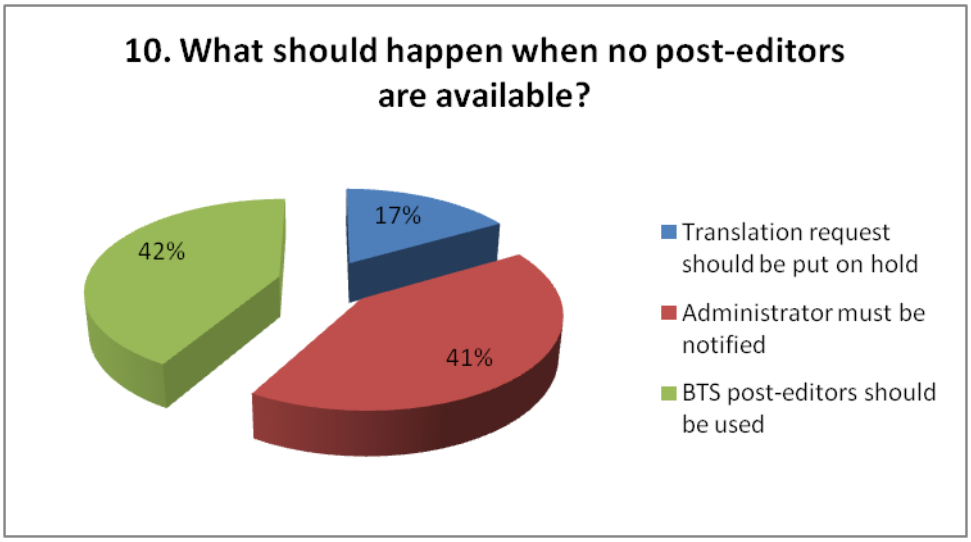

5. What type of post-editors would you typically use? (multiple answers allowed)
80%
70%
60%
50%
40%
30%
20%
10%
0%
75%
58%
25%
8%
33%
Translation requesters will post-edit
Other internal resources will post-edit
Dedicated external party will post-edit
Ad-hoc external party will post-edit
BTS service should provide post-editors

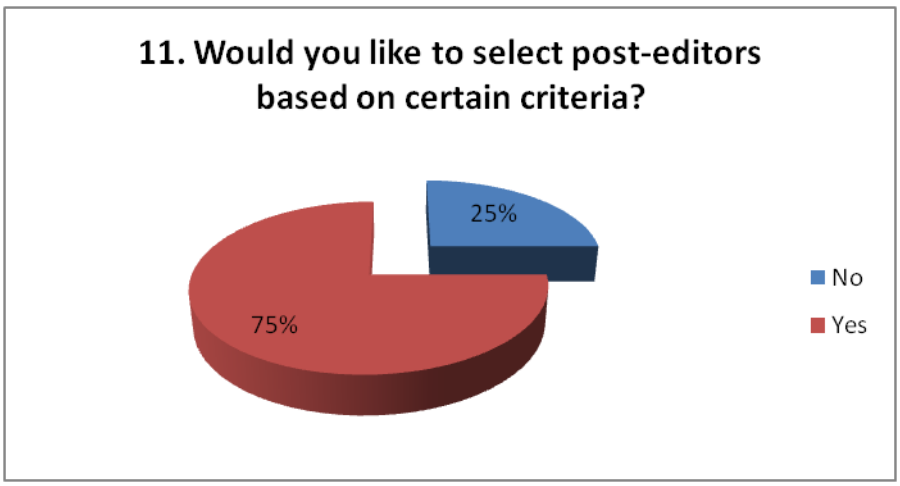

6. How many post-editors do you expect to use on the Bologna Translation Service?
17%
8%
8%
67%
1
2-5
6-10
More than 10

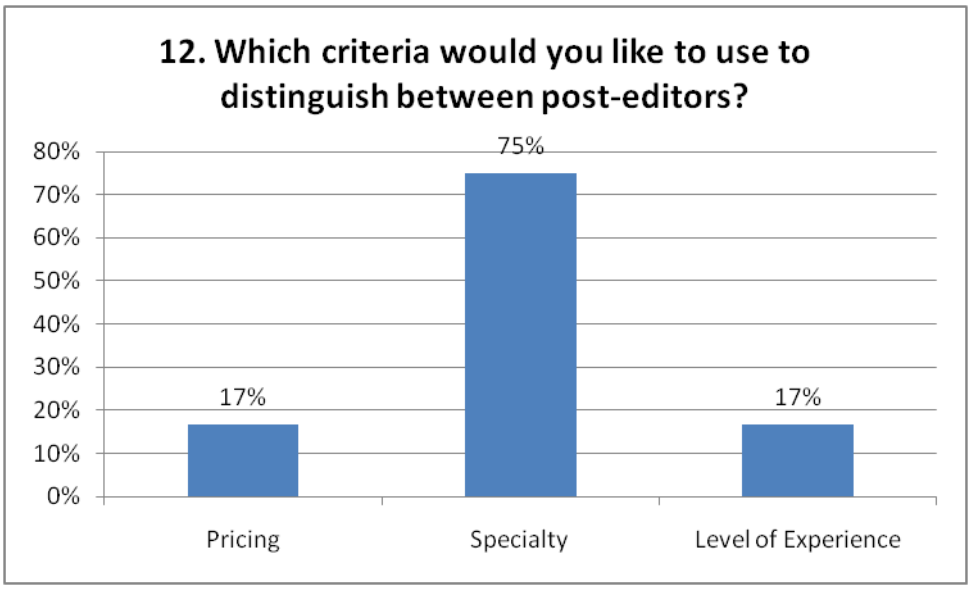

7. How would you like to assign a post-editor to a translation request? (multiple answers allowed)
70%
60%
50%
40%
30%
20%
10%
0%
25%
25%
25%
58%
Requester assigns post-editor when making request
Post-editor is assigned automatically
Post-editor is assigned automatically but may be changed by requester
Company or university administrator assigns post-editor

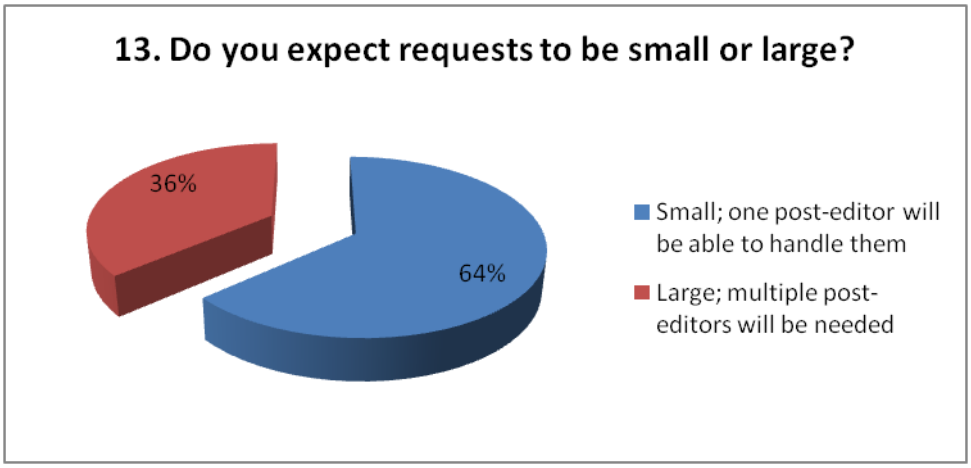

8. Should a post-editor be able to reject an assigned post-editing job?
30%
70%
No
Yes

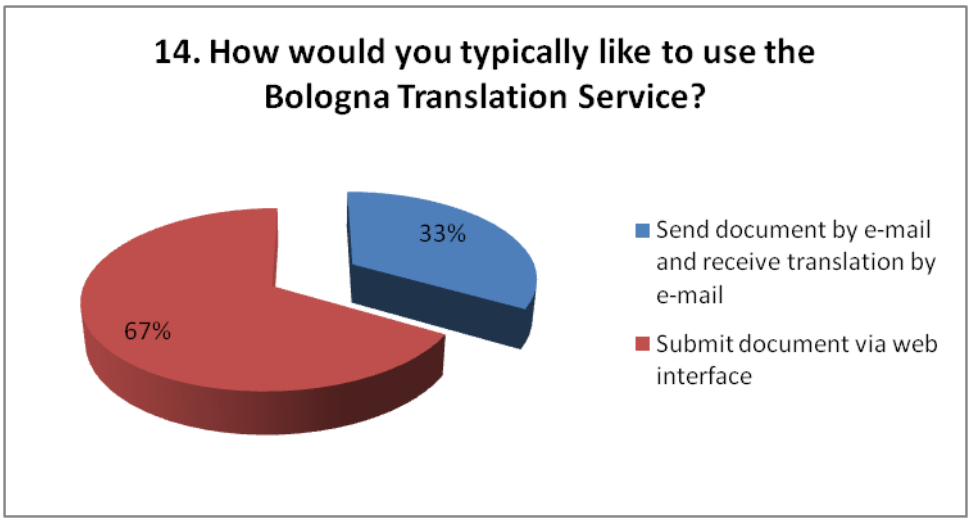

9. What should happen if post-editor rejects a job?
25%
25%
12%
38%
New post-editor is assigned automatically
Rejecting post-editor assigns new post-editor
Company/university admin selects new post-editor
Requester selects new post-editor

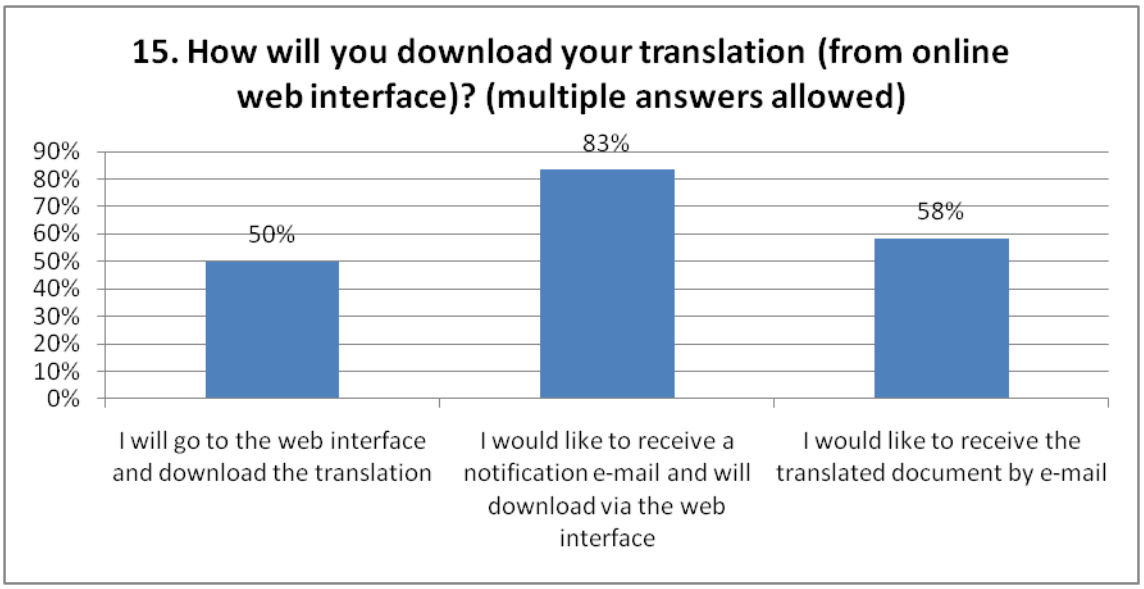

10. What should happen when no post-editors are available?
17%
42%
41%
Translation request should be put on hold
Administrator must be notified
BTS post-editors should be used

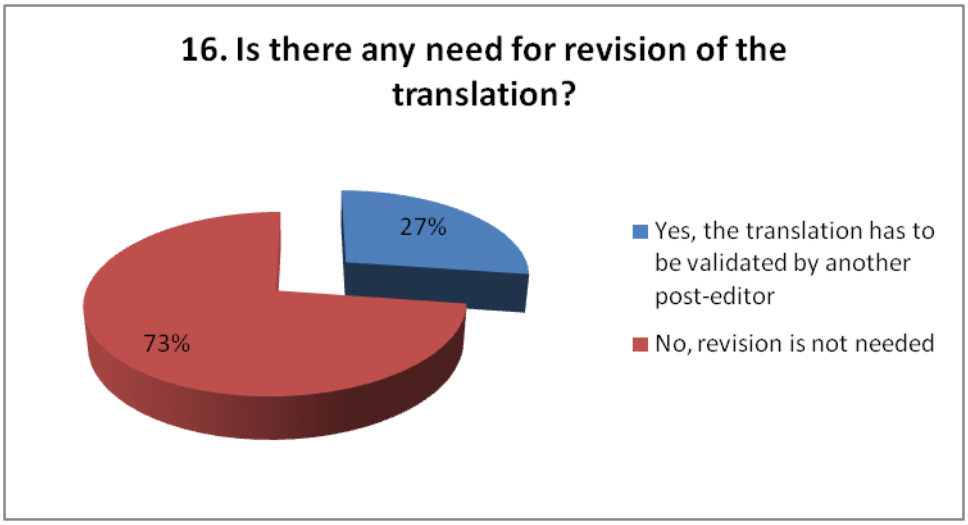

11. Would you like to select post-editors based on certain criteria?
25%
75%
No
Yes

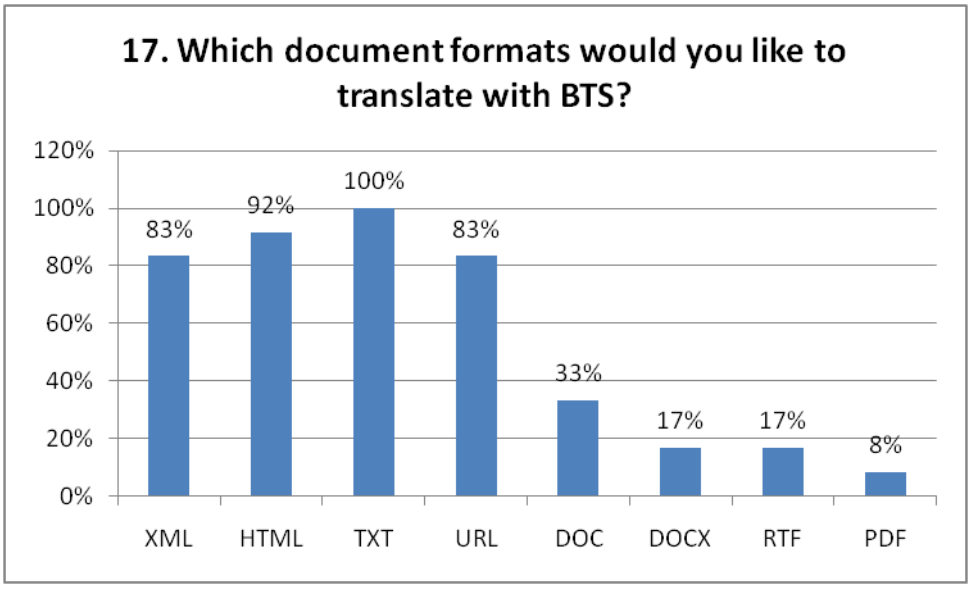

12. Which criteria would you like to use to distinguish between post-editors?
80%
70%
60%
50%
40%
30%
20%
10%
0%
17%
75%
17%
Pricing
Specialty
Level of Experience


13. Do you expect requests to be small or large?
36%
64%
Small; one post-editor will be able to handle them
Large; multiple post-editors will be needed


14. How would you typically like to use the Bologna Translation Service?
33%
67%
Send document by e-mail and receive translation by e-mail
Submit document via web interface


15. How will you download your translation (from online web interface)? (multiple answers allowed)
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
50%
83%
58%
I will go to the web interface and download the translation
I would like to receive a notification e-mail and will download via the web interface
I would like to receive the translated document by e-mail


16. Is there any need for revision of the translation?
27%
73%
Yes, the translation has to be validated by another post-editor
No, revision is not needed


17. Which document formats would you like to translate with BTS?
120%
100%
80%
60%
40%
20%
0%
83%
92%
100%
83%
33%
17%
17%
8%
XML
HTML
TXT
URL
DOC
DOCX
RTF
PDF