Overview of PwC/Sytranet on-line MT Facility

Ross Smith
Translation Service
Company Administration Services
PricewaterhouseCoopers
Madrid
Spain
ross.smith@es.pwc.com

This paper describes the on-line machine translation service implemented by PricewaterhouseCoopers (PwC) in early 2000 on its Intranet (called KnowledgeCurve) using SYSTRAN MT technology.

The paper is divided into three sections. The first gives a description of the reasons for setting up this system, while the second provides user statistics and feedback, in addition to practical examples of how the MT engine is used to obtain a working translation into a language the user knows but in which he is not fully proficient, or for gisting a document in an unknown language.

The third section concerns potential improvements to the system in the future. In particular, reference is made to the translation of financial reporting documents using Extensible Business Reporting Language (XBRL), taking advantage of another PwC-SYSTRAN joint project which is pioneering the use of XBRL in a web-based service for translating financial documents, and a further XBRL initiative undertaken by PwC with Nasdaq and Microsoft.

1. Background: reasons for implementation

Following the world-wide merger between the professional services firms Price Waterhouse and Coopers and Lybrand, a Knowledge Management team headed by senior consultant Jonathan Sage was set up in 1999 to investigate translation services around the world and possibly identify an international supplier of language services to offices in all countries. This was no small aim, since PwC is present in 142 countries, with over 125,000 professionals.

During the initial stage of researching the needs of offices around Europe, it was discovered that either individual translators (ranging from secretaries doing occasional translation work to fully qualified professionals) or translation departments already existed in a number of major offices, particularly Paris, Amsterdam and Madrid. Finding a global translation service supplier was therefore no longer relevant, and in fact would probably have been opposed by users of translation services who we may assume were reasonably happy with the service they were receiving locally.

Instead, the decision was taken to go ahead with the original concept of a global translation service but via a different channel, making basic translation available to all PwC employees in any location through access to a machine translation engine on the

firm's Intranet. The Spanish firm, which had an internal translation department and a strong Knowledge Management team, was closely involved in the project from the outset.

It was made clear to the Knowledge Management team by the experienced translators working at the Madrid office and elsewhere that the machine translation programs available at the time would not be capable of translating to an acceptable standard the kind of documents that were typically handled by the PwC translation departments (annual accounts, financial reports, international taxation analyses, commercial contracts, promotional materials, etc.) because of their highly technical nature and the complex language involved. MT systems are good at translating weather reports, user guides or help files for electronic consumer products, basically any document with simple syntax and restricted vocabulary. They are not good at translating grammatically complex analyses of different areas of law or subjective advice on how best to minimise the tax burden of foreign investments, to give just two examples. And in any case, the one area where computers were definitely of help for PwC's translators, this being the use of CAT for documents with high levels of repetition (essentially, annual accounts and certain contracts), was already covered, or would soon be covered, in the major offices through the use of Trados translation memory software.

No attempt was made, therefore, to give the impression that MT would provide a universal solution for PwC's translation needs; rather, it was presented as an alternative to "human" translation for certain circumstances in which high linguistic quality was not an essential prerequisite and fast, dynamic translations were required.

The largest industry players were approached and SYSTRAN was chosen from among the leading suppliers as they could combine state-of-the-art technology with the largest breadth of available language pairs. In addition, SYSTRAN's online translation service is a turnkey solution hosted on SYSTRAN's servers. This efficient approach allowed PwC's service to be up and running in a very short timeframe without tapping into the company's IT resources. SYSTRAN were already known to the team from PwC since they provided the basis for the European Commission MT service and also the engine for the popular Babelfish service on the Alta Vista browser.

Trials with machine-translated documents were carried out by the Knowledge Management centre with a target group of volunteers in various countries and the site commenced operation on a pilot basis in the Spanish firm in early 2000. Soon after, the on-line service was made available to PwC employees in all countries.

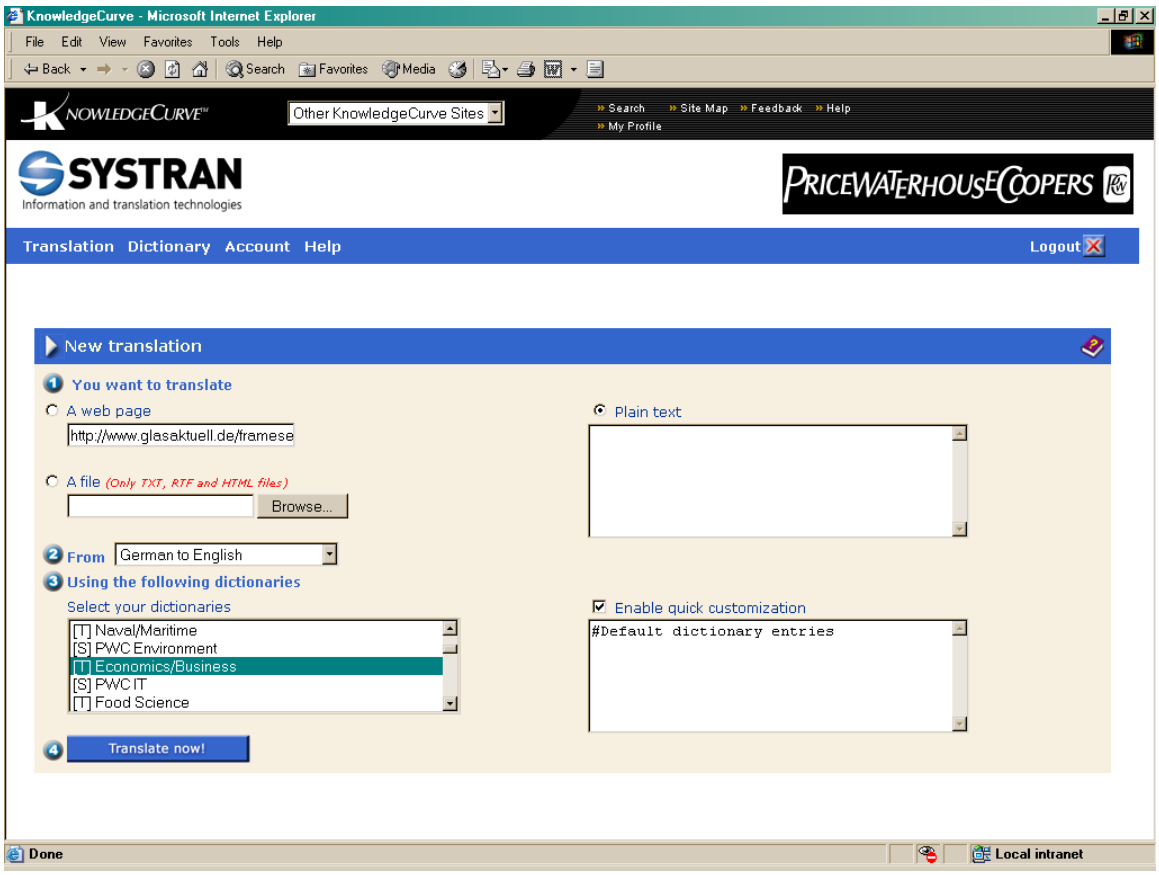

Figure 1: User interface of PwC/SYSTRAN web site

The translation engine currently offers 37 language pairs and 21 SYSTRAN dictionaries on specific subject areas.

During 200l and 2002 specific dictionaries provided by PwC were incorporated into the system in order to help tailor it, to the extent possible, to the firm's terminology. A number of English-Spanish dictionaries on areas such as insurance, banking, tax and auditing, migrated from Trados MultiTerm, were provided by the Madrid Translation Service, and English-French and Japanese-English dictionaries were added later on XBRL. Users can choose between these dictionaries and the ones supplied by SYSTRAN.

In addition to these specialised dictionaries, PwC employees have the option of creating their own dictionaries, in which they introduce the terms themselves. The procedure for doing this is set out in considerable detail in an on-line user guide. To date, 216 such user dictionaries have been created which are only available to the user that made them. This reflects a substantial interest in the facility on the part of these users.

In this connection, it should be stressed that SYSTRAN also offers its customers a range of professional services for high-end customisation. The service utilised by PwC has not been customised in depth in order to maximise the quality of the output since this was not the overriding purpose envisaged for the system by PwC when it was started up.

PwC employees around the world have access to the firm's intranet, called KnowledgeCurve, which provides them with all sorts of information on the firm itself, its lines of service, clients, markets, etc., in addition to internal services and data specific to employees in the various countries.

As shown below, the PwC/SYSTRAN site is accessed on the KnowledgeCurve home page through a list of Quick Links to useful sites or services on the intranet. The Spanish firm's site also contains a direct link from its main page.

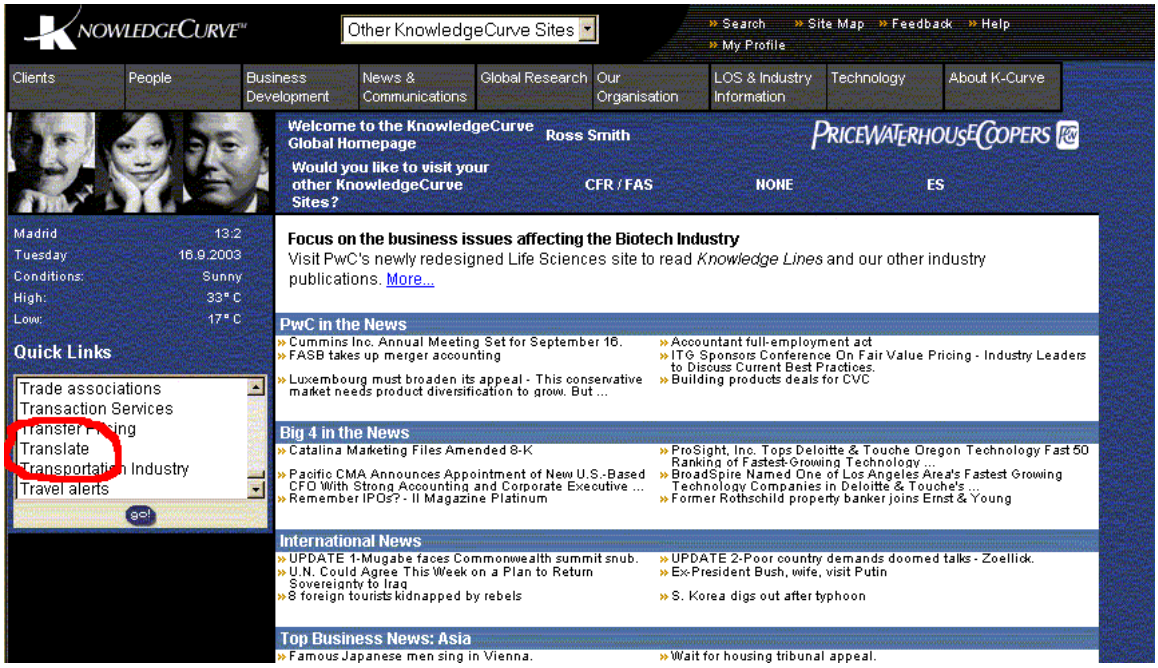

Figure 2: KnowledgeCurve home page showing translation link

Users have to introduce a user name and password to access Knowledgecurve, and this same password is valid for the MT site. This avoids the need to write in the user ID and password twice. Employees can also access the site from the Internet, rather than via the intranet, in which case their PwC email address serves as the user ID.

2. MT in action: what users think, and why they use the on-line facility

This section looks at how, why and by whom the PwC/SYSTRAN MT facility is used, providing data on most popular language combinations, users' opinions on the worth of the system, and three practical examples of how the facility is actually utilised by PwC professionals.

a) Most requested combinations

The most popular language combinations of the 37 available pairs, with the number of translations requested, are shown in the following table (September 2003):

Language Pairs	No. of requests
English-Spanish	40327
Spanish-English	22727
German-English	16756
English-German	8567
French-English	7304
English-French	5278
Dutch-English	4764
Portuguese-English	1942
English-Portuguese	1541
English-Chinese (simplified)	1485
Italian-English	1396
English-Italian	1360
Japanese-English	1273
English-Dutch	1271
English-Chinese (traditional)	1244
German-French	1015
French-Spanish	920
Dutch-French	818
French-German	697
French-Dutch	692
German-French	646

Table 1: Most requested language combinations

The most requested language pairs for translations in both directions (e.g. English-Spanish and Spanish-English) are reflected in the following graph. The figures for traditional and simplified Chinese have been merged.

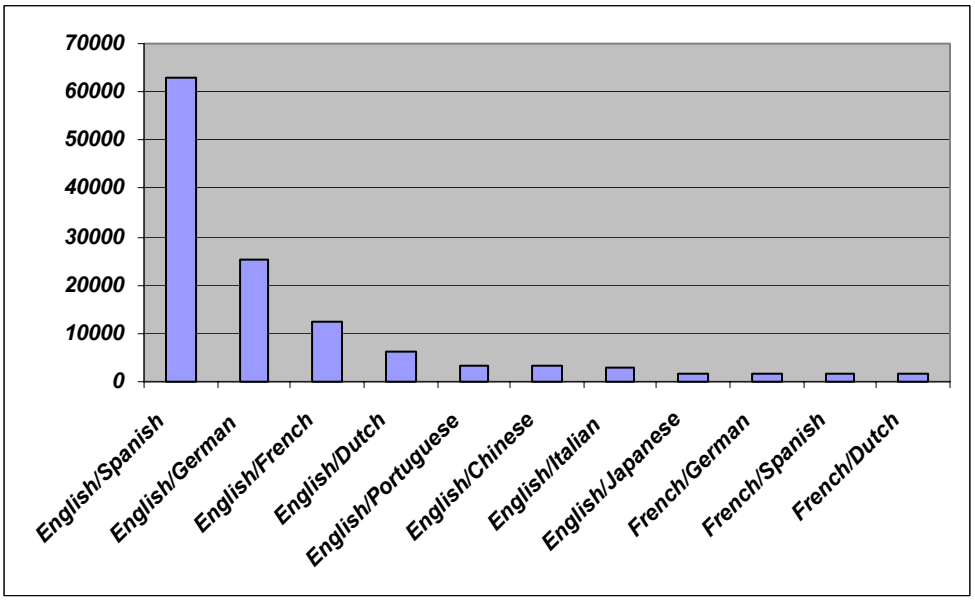

Figure 3: Most requested language combinations

The total number of translations requested for all languages is around 130,000 and some 7,300 PwC employees around the world have used the system. As can be seen from the above data, combined requests for English-Spanish and Spanish-English account for almost half of the total for all languages, and more than double the next most popular combination, German-English-German. This backs up the widely held view in the language community that Spanish is the world's most international language after English.

b) User feedback

The PwC/SYSTRAN contains a mechanism, called "linguistic feedback", whereby users are able to provide information on their use of the facility and their opinions on the results obtained. In addition to data on the type of document (text, file, web site) and language combination used, the feedback form asks users to comment on strengths or weaknesses observed by them and to indicate whether they agree with the following statements:

[ ] I could basically understand the translation
[ ] I found the translation suitable for my requirements
[ ] I am still favourable to using this technology to help in my comprehension needs on FL text

With regard to the responses received from users, both directly from the intranet site and via a survey of professionals in the Spanish firm using the same format, 57% had ticked

at least one of the boxes and may therefore be regarded as “positive”, while 43% had no box ticked, and should therefore be considered “negative”.

Of the positive responses, opinions concerning the value of the facility (i.e. boxes ticked) can be analysed as follows:

Response	% of respondents
Could basically understand translation	54%
Found translation suitable for requirements	21%
Still favourable to using this technology	25%

Table 2: User feedback opinions

Graphically, this may be represented as follows:

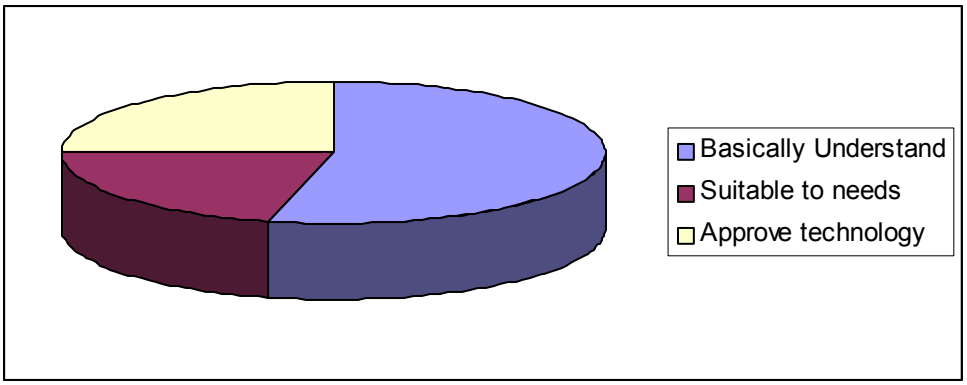

Figure 4: User feedback opinions

Concerning the format options for the original documents (web address, file or text), the vast majority of users (80%) chose the text option, either writing or pasting their source document in the pertinent space. A file was uploaded for translation by 15%, while 5% requested the direct translation of a web site.

Specific comments from users reveal that the PwC/SYSTRAN facility is used basically to obtain a rough idea of a document’s content, to prepare drafts in other languages and even to translate single terms. Opinions as to its value vary considerably, as can be appreciated in the verbatim selection of user comments given below, in descending order of enthusiasm:

“I tried it out and think it’s amazing!”
“I consider this option to be very useful”
“I reckon this is a useful tool”
“For simple translations it seems pretty good”
“It’s useful for certain texts, but the translations are not very accurate”
“I think it could be useful for translating or enquiring about single words, but with regard to texts in general the translation is incongruent”
“I wouldn't use this machine for more complicated texts, because it's not reliable”
“Translation failed utterly”

This last, wholly negative response is not surprising, since the user was trying to translate into English from Dutch using the German-English engine!

c) Practical examples

Users of the PwC/SYSTRAN who accessed the “Language” page on the PwC intranet during the early stages of the facility’s development were warned about the limitations of MT and advised not to use it for complex or colloquial texts. The more negative comments cited above are doubtless from employees that did not heed this advice. Communications on the subject from the Madrid Translation Service stress that one of MT’s main advantages is allowing users to gist foreign language texts to obtain a general idea of their content before deciding whether or not to request an accurate “human” translation, while another is to help draft documents in languages of which the user has some knowledge but is not fully proficient. It is not a replacement for professional “human” translation.

There follow three “real life” examples of how the MT facility has been used by the Translation Service at PwC Madrid.

The first instance concerns the translation from German to English of a web site at the request of PwC professionals in the legal department who wished to obtain a basic idea of its content before deciding whether to commission a formal translation.

The second example consists of the translation from German to English of a Word document so that the translation manager could ascertain the subject matter of the document and on that basis decide which translator it should be assigned to.

The third case concerns the translation of a letter to a client from Spanish to French, where the grammatical accuracy of the final product was not of key importance, and there was a desire to avoid the cost of sending such a short text to a professional translator.

Example one: German-English translation of Web site

A lawyer from PwC's tax and legal practice asked the Translation Service for a rough translation of a German language web site in relation to a case she was handling, to find out what it was about and whether a full translation was needed.

No-one in the internal translation team speaks German. Therefore, the URL of the site was pasted directly into the "translate web site" option on the PwC/SYSTRAN screen and the output obtained was (partially) as follows:

Did you know that glass was not dedicated ever one day? There is the year for threatened birds, the month of the tree or the father day.One gives its attention to all possible one.But one looks for one day of the "glass" in vain.A little actually amazingly, one considers what important role glass in our life plays.For us reason of enough to dedicate to this packing an InterNet side.

What can you find on this side?
Now, in whole simply, everything approximately around glass.
They find here useful, unterhaltsames, strange, interesting and - each quantity beautiful glass containers. Much fun.
The action forum exists since at the beginning of 2000 and offers a platform of the specialized combination container glass industry e.V. and the GGA company and waste avoidance ltd..
With this action forum one wants to react to the constantly changing market. The action forum glass packing understands container glass as a system, with which production, use and recycling form an inseparable unity.

Sample translation 1: Web site, German-English

The translated text may contain a number of errors, but the overall content is quite clear and one has no trouble ascertaining what the site is about. Precisely why a German language site praising the benefits of glass as a material for making containers should be of interest to a member of our legal department, it was not our job to ask.

Example 2: Word document in German

Again, as there is no in-house German translator at PwC Madrid, the MT intranet facility was used to obtain a rough understanding of the content of a document sent for translation from German to Spanish. The Translation Service uses a number of external translators who have different areas of specialisation. Before deciding to whom the translation should be sent, therefore, it was desirable to know the document's subject matter (whether it was financial, accounting, legal, technical etc.) since we had no references whatsoever in this regard.

The output produced by the PwC/SYSTRAN site was as follows:

```
The order was given to us to draw up a "land overview taxes" for Spain.

The "land overview taxes" is a compressed representation of the fiscal
basic conditions for Spanish real estate investments of a property
special estate.

This representation is to serve the department competent for
international investments as basis for real estate acquisitions and
cover the following points:

· possible investment structures
· taxation of the individual investment structures in the foreign
country and inland during the individual project phases (purchase,
letting and sale)
. In particular thereby the following aspects are to be treated:
· direct investment versus investment over a foreign real estate
company
· ate Deal versus Share Deal.
· optimization of purchase of land expensive and/or value added tax
with the purchase
· numerical representation of the tax burden.

The land overview is not a comprehensive representation of the fiscal
basic conditions for real estate acquisitions.
```

Sample translation 2: Unknown document, German-English

This is how (part of) the “raw output” looked on screen, together with the source text in German:

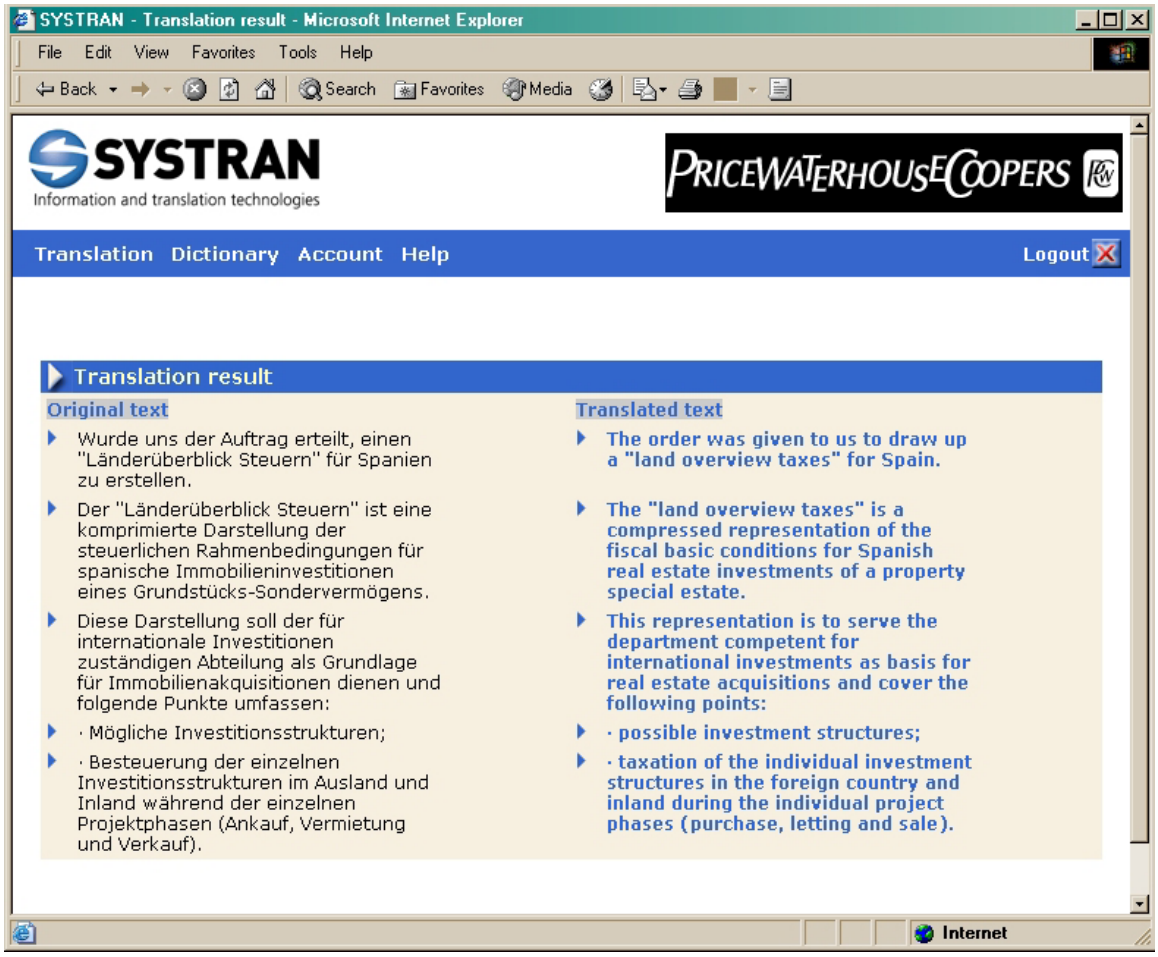

Figure 5: Result of translation request as it appears on screen

The translation contains errors and the English is not very natural, but it is still easy to ascertain that the document concerns a request for advice on the tax treatment of real estate acquisitions and the different investment structures that could be implemented. It was therefore sent for translation to a legal specialist with knowledge of tax matters.

Example 3: Spanish to French translation of letter to client

In the third example, the Translation Service was asked to handle a short translation from Spanish into French. The translator concerned, who is capable of translating accurately from French but is not so confident into French, used the PwC/SYSTRAN facility output as a basis for the French text, which was then put through the MS Word French spell check and grammar check. If there was doubt about the Spanish-French output, the translator tried translating the English equivalent into French instead. The final document was deemed good enough to send to the client (and no negative feedback was received thereafter).

The source document and translation are given below:

Spanish original:

Apreciado Sr.Chirac:
De acuerdo con su aprobación y, siguiendo las instrucciones de la Sra. Anne Brown, adjunto le remito los originales de las Cuentas Anuales de la sociedad "Mercantil España, S.L , sociedad unipersonal" del ejercicio finalizado el 31 de Diciembre de 2.000.
Según indicaciones de la Sra Brown en su e-mail del 26 de septiembre, Sra.Anne Henry se encargará de que se firmen y de remitirlas a Cabinet DESCARTES & ASSOCIES para que sean depositadas en el Registro Mercantil.
Si precisa cualquier aclaración, no dude ponerse en contacto con nosotros,

Atentamente,

French translation:

Cher M Chirac :
Selon votre approbation, et après les instructions de Mme Anne Brown, nous joignons les originaux des comptes annuels de la compagnie "Mercantil España, S.L, sociedad unipersonal" concernant l'exercice terminé le 31 décembre 2000.
Selon les indications de Mme Brown dans son E-mail de 26 septembre, Mme Marie Henry se chargera de les faire signer et de les délivrer à Cabinet DESCARTES & ASSOCIES afin qu'ils soient déposés au Registre Mercantile.
Nous restons à votre disposition pour tout renseignement supplémentaire que vous souhaiteriez recevoir.

Avec nos sentiments les meilleurs.

Sample translation 3: Letter to client, Spanish-French

This is evidently a rather laborious way of obtaining a 92-word translation, but it is an interesting reflection of how a combination of commercial tools (SYSTRAN and MS Word) can be used to achieve an acceptable result.

C) Conclusions

Condensing all the above information, it is clear that the on-line MT service is popular among PwC employees working in a multilingual environment. Over 7,000 staff members have used the system since it was introduced, with around 130,000 translation requests. As is usually the case with MT systems, users' opinions vary a great deal, from open enthusiasm to frank criticism.

The reaction of first-time users depends on a number of factors which vary greatly from one person to another and which explain the disparity in levels of satisfaction. These factors are: users' initial expectations based on the extent of their knowledge of similar systems and their faith in IT in general; the extent of their knowledge of the language into which they wish to translate; the language combination they choose (this is difficult to evaluate subjectively, but it seems clear that MT systems handle certain languages better than others); the subject matter of the source document, and the grammatical quality of the source document.

Most users are not linguists themselves: they feed the documents in and judge the results without taking any of these mitigating factors into account, and express their opinions accordingly.

Although a number of users have voiced their scepticism as to the system's usefulness, a large percentage seem happy with the service they receive and a majority of them consider that at least they were able obtain a basic understanding of the document they wished to translate.

The vast majority use the system for translating short texts, which are written or pasted directly into the "plain text" box on the site, or even single words. Spanish/English is the most popular language combination, followed at considerable distance by French/English and German/English. The most popular language pair not involving English is German/French. It is interesting to note the relatively high number of requests for English/Chinese translations (roughly the same as English/Portuguese and above English/Italian), reflecting the increasing importance of China in the business world.

The MT facility is considered most useful for gisting documents in an unknown language, seeking translations of individual terms and preparing draft documents in a language known by the user, but in which he/she does not feel fully competent (usually English), which may subsequently be checked using dictionaries and spelling and grammar correction functions in word-processors (e.g. MS Word). Users in general are aware that the facility is not capable of producing "human" quality translators, and therefore do not use it for that purpose.

3. Future enhancement: XBRL

PwC/SYSTRAN XBRL on-line translation facility

This project, involving the same partners as the service described above, is still at the prototype stage and is ultimately intended for commercial, rather than in-house, utilisation.

PricewaterhouseCoopers has teamed up with SYSTRAN to develop a web-based facility which unites multilingual technologies with XBRL. Based on Extensible Markup Language or XML, XBRL streamlines the way companies report and publish their financial data, and how analysts and investors can review that information.

XML and its various special-purpose offshoots, which include XBRL, has been highly successful and has been adopted on a large scale as the new standard mark-up language for web documents. It offers a high degree of reliability when composing and modifying documents and at the same time is an open system which can be used on a wide variety of platforms.

As the Internet is no longer dominated by the English language, the demand for multilingual content is being accelerated. SYSTRAN's technology adds a multilingual layer to the automated process that allows for on-demand and timely publishing of financial information in various languages.

The PwC/SYSTRAN XBRL facility intends to take advantage of the fact that numerous financial reporting documents for web use are being written in XBRL to provide a service offering exact translations of financial terms. A prototype version for Japanese is shown below.

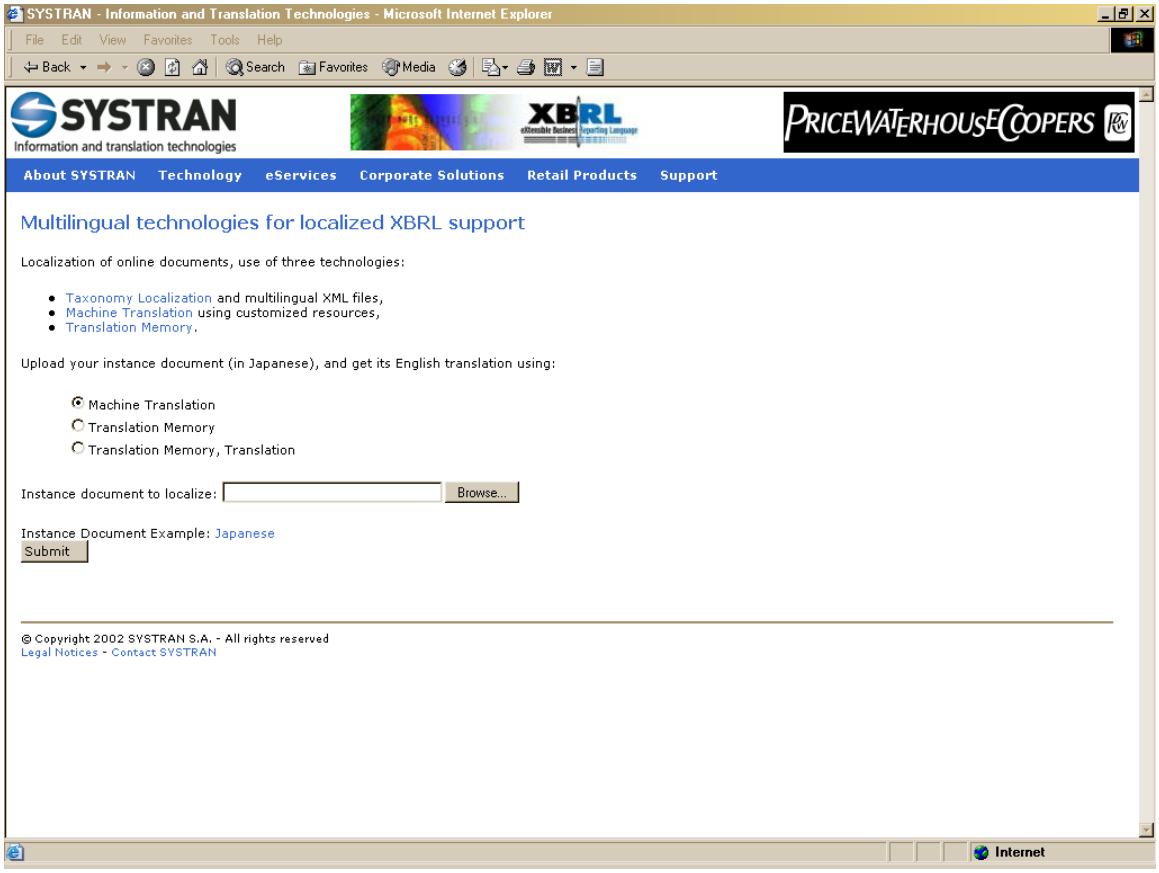

Figure 6: User interface for PwC/SYSTRAN XBRL web site

Using this technology, a person who does not speak the language in which financial documents are created will be able to search the Internet for the information from a company and be presented with the data automatically in the language of his choice. This technology provides major opportunities for the publishing and analysis of corporate reports around the world. The system would be particularly useful for investors and financial analysts needing data on companies that place their financial information on the Internet in languages not understood by these users. The multilingual standardisation of the formats used for reporting financial data will promote corporate transparency, accountability and integrity on a truly global basis.

The advantage of XBRL for MT, in addition to its reliable and straightforward structure, is that the codes which are already embedded in the XML or XBRL document can be utilised in the MT process. Apart from specific terms or phrases (“taxonomy elements”), codes are also included which indicate the language in which the document has been prepared, which is logically helpful for the MT programmers.

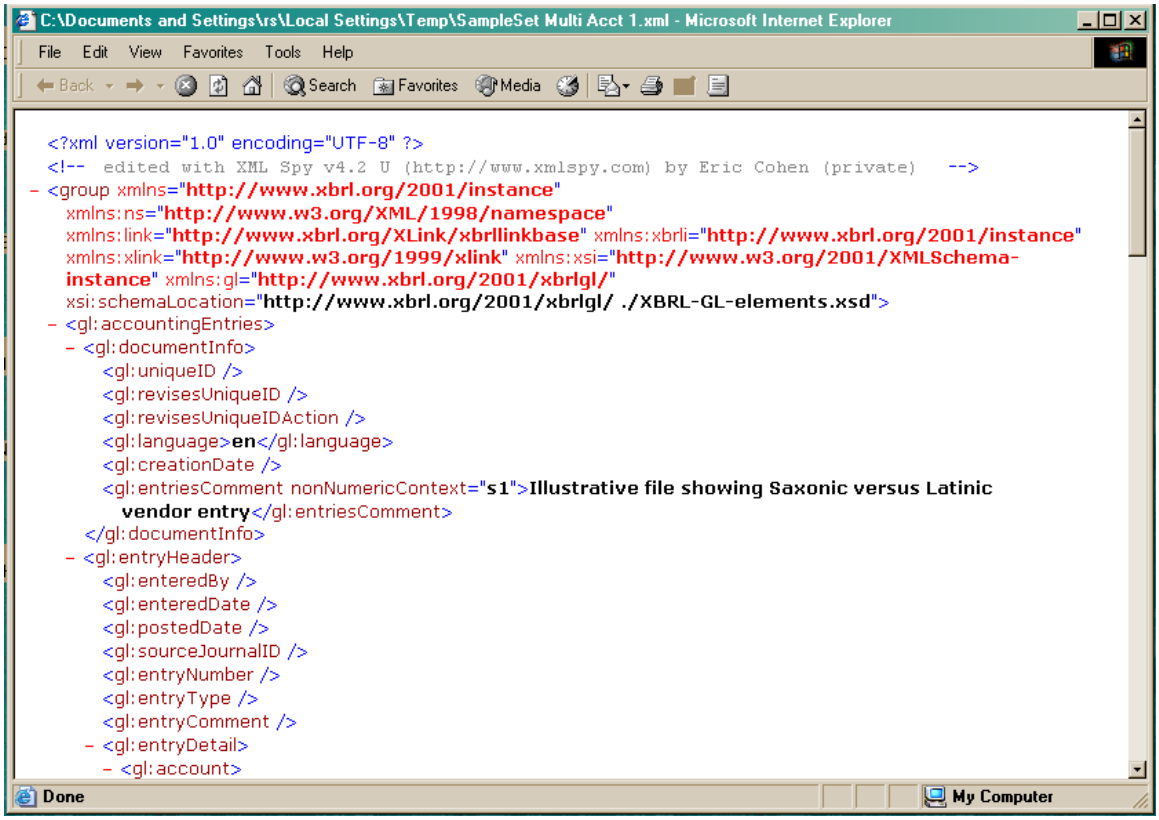

Figure 7: sample XBRL document showing language tag

XBRL project with Nasdaq and Microsoft

PwC is also working with Nasdaq and Microsoft on an Internet service, based on Microsoft's Excel spreadsheet, whereby users will be able to combine both stock market data provided in XML and financial data tagged in XBRL to generate documents with minimum difficulty. The terms used in the XBRL documents are taken from International Financial Reporting Standards (IFRS), which provide a highly standardised set of financial reporting terms that can be used as a basis for mapping similar standardised terminology in other languages to create very accurate translations of financial documents. For instance, if the English term "Property, plant and equipment" has a specific tag so that whatever the IT platform used to manipulate or display the XBRL document it will always appear as "Property, plant and equipment" with the same potential set of subheadings (equipment, buildings, vehicles, etc.), it should be straightforward to substitute in the relevant equivalents in other languages taking advantage of the existing tags. This is potentially a powerful MT tool.

Once the PwC/SYSTRAN XBRL service is operating regularly it could be added to the currently available range of options (translating a website, uploaded document, plain text, etc.) on the PwC/SYSTRAN on-line MT facility on KnowledgeCurve to provide users with the chance to obtain very accurate translations of financial reporting documents.

The range of document types involved would logically be restricted to purely financial texts, such as profit and loss accounts, balance sheets, cash-flow statements and notes to annual accounts using internationally accepted accounting terms in English, with recognised equivalents in other languages . However, the advantage would be that, in principle, almost total accuracy could be achieved due precisely to the very limited vocabulary involved. It is also worth taking into account that PwC employees handle just this sort of documents in their day-to-day work with international clients.

PwC will continue to work with SYSTRAN and its other business partners to find ways of improving the existing on-line MT services, for the benefit of both the firm's employees and its clients.