[*Translating and the computer 12: Applying Technology to the Translation Process*. Proceedings of a conference … 8-9 November 1990, ed. Catriona Picken (London: Aslib, 1991)]

Language technology products in the European Market

Brigitte Engelien

Ovum Ltd

First, a few words about Ovum: we are a London-based consultancy company which specialises in studies on new technologies in computing and communications. The company is now five years old, and there are 40 of us. The very first of our published market research reports, back in 1985, was about natural language computing, and we have been keeping an eye on developments in that area ever since. Our most recent knowledge in this area stems from a study for the European Commission which we are just completing, together with the Paris consultancy SIAR Bossard, whose object it was to find out:

— just how many people in Europe spend how much time doing the kinds of things which natural language computer systems can help them do better or faster; and
— what language technology products are actually on offer out there, how real are they, and what are their chances of commercial success.

These were the main components of this study:

— a survey of 1,100 office workers throughout Europe (who are the main target group for language technology products, although there are other potential users);
— a programme of interviews with suppliers of language technology products;
— and, as you always get a glowing picture of their products from suppliers, a number of interviews with users;

— from these three information sources, we put together some tentative scenarios as to how the market might develop, and invited a Round Table of experts to comment on them;
— finally, we produced, based on their input, forecasts of users and sales to the year 2000.

Language processing technology has been applied to a wide range of application areas, not just to foreign language translation. The problems of parsing and semantic analysis are always there, but different approaches have been taken to solve different problem situations.

Almost all language technology products actually perform translation of one kind or another: natural language interfaces to relational databases translate the user's natural language query into the appropriate computer command language (usually SQL) to extract the desired information; message scanning systems translate a piece of free-form text into a message of a clearly defined format; and talkwriters translate speech into text. The different products complement each other to some extent, and it is easy to see how with time, several will be combined to form more powerful tools.

My intention today is to give you the broader picture of language technology products today. First, some selective results from the user survey. Next, an overview of language technology products today and how they are expected to develop; and last, to make it all more tangible, I want to tell you a bit more about one particular LT product: the talkwriter.

There are 45 million office workers in Europe (or rather, in the 12 EC countries; I feel quite un-European at this conference with its Eastern European flavour, leaving out substantial chunks of the new Europe). Between them they spend 9 million person-years per year - or just under a fifth of their time - dealing with text of some sort.

Figure 1 shows the activities on which their time goes, in decreasing order of person-hours:

— copy-typing
— reading documents received
— writing
— editing
— searching for information
— reading information retrieved
— consulting management information
— filing documents or information
— reading in a foreign language
— writing in a foreign language
— translating.

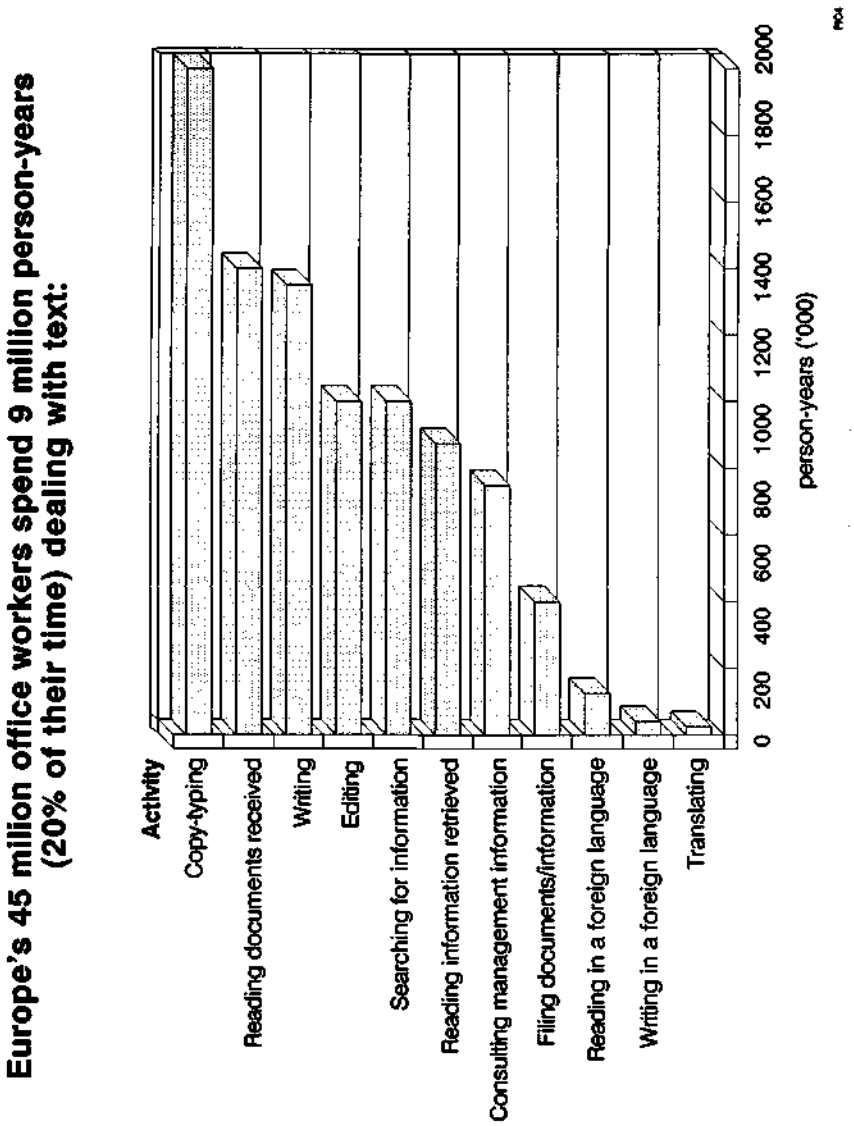

Europe's 45 million office workers spend 9 million person-years
(20% of their time) dealing with text:
Activity
Copy-typing
Reading documents received
Writing
Editing
Searching for information
Reading information retrieved
Consulting management information
Filing documents/information
Reading in a foreign language
Writing in a foreign language
Translating
0 200 400 600 800 1000 1200 1400 1600 1800 2000
person-years ('000)

Figure 1

The last three activities are the smallest in terms of person-hours, not surprising really. This is only relative, however – the total numbers of people hiding behind these numbers are still very large.

This gives you a general picture of the survey results; the next three figures are concerned with foreign language activity only, because I thought that would interest you most.

In order to arrive at potential users for the different LT tools what matters is *how many* people perform these activities – and how many perform them frequently enough to contemplate the use of an LT tool.

In Figure 2, the office workers are broken down by professional layer. The survey was organised around five separate layers:

— self-employed professionals
— salaried professionals
— senior management
— middle management
— other office workers.

The shaded area shows the percentage of people in each group who read foreign language material at least once a month. The rate is really quite high in all groups – 25% or more – but highest amongst senior management. So the demand is there; the question suggests itself: how many more people would read foreign language material if they were able to?

Figure 3 shows the people (in the same professional layers, and still all across the EC) who produce written translations at least once a month. There are six million people who do this, the highest proportion of them amongst the middle management layer.

People who produce written translation daily were too few (relatively speaking) to show up in the graph; I have listed them over to the right, and they still add up to 1.3 million people across all the layers.

Please don't be put off by how complex Figure 4 looks; the idea is quite simple: it relates to all the foreign language activities together (reading, writing, translating). Starting on the left, there is a section relating to the same professional layers we have seen already. This is 100 per cent of the office population; to the left of it is 100 per cent of the foreign language activity person-hours spent in a year. So this shows that self-employed professionals, who make up 7 per cent of the office population, are responsible for only 3 per cent of the total person-hours spent on foreign language activities; whereas the senior management group, which makes up 10 per cent of the office population, is responsible for 29 per cent of foreign language activity.

The middle section shows the same idea, but in relation to industry sectors. This is 100 per cent of the office population, broken down into relevant industry sectors: manufacturing industry, banking and insurance, administration, and other services. You can see that manufacturing

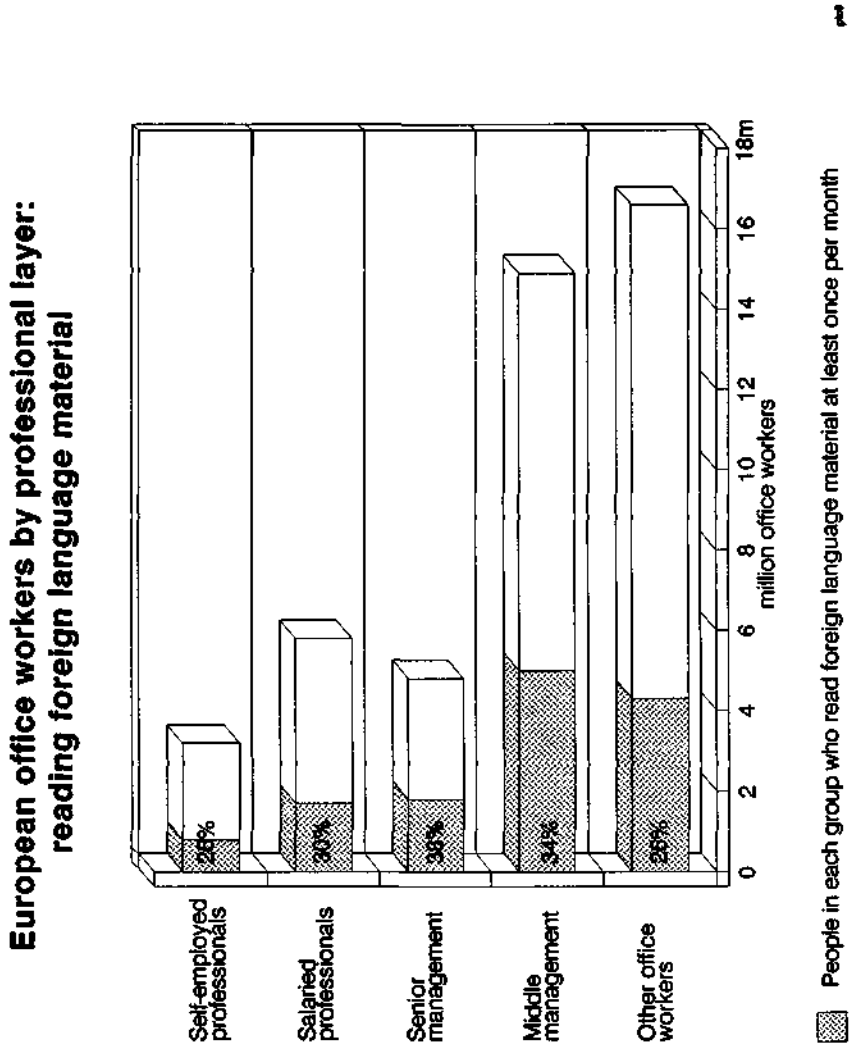

European office workers by professional layer:
reading foreign language material
Self-employed professionals
Salaried professionals
Senior management
Middle management
Other office workers
28%
30%
38%
34%
25%
0
2
4
6
8
10
12
14
16
18m
million office workers
People in each group who read foreign language material at least once per month

Figure 2

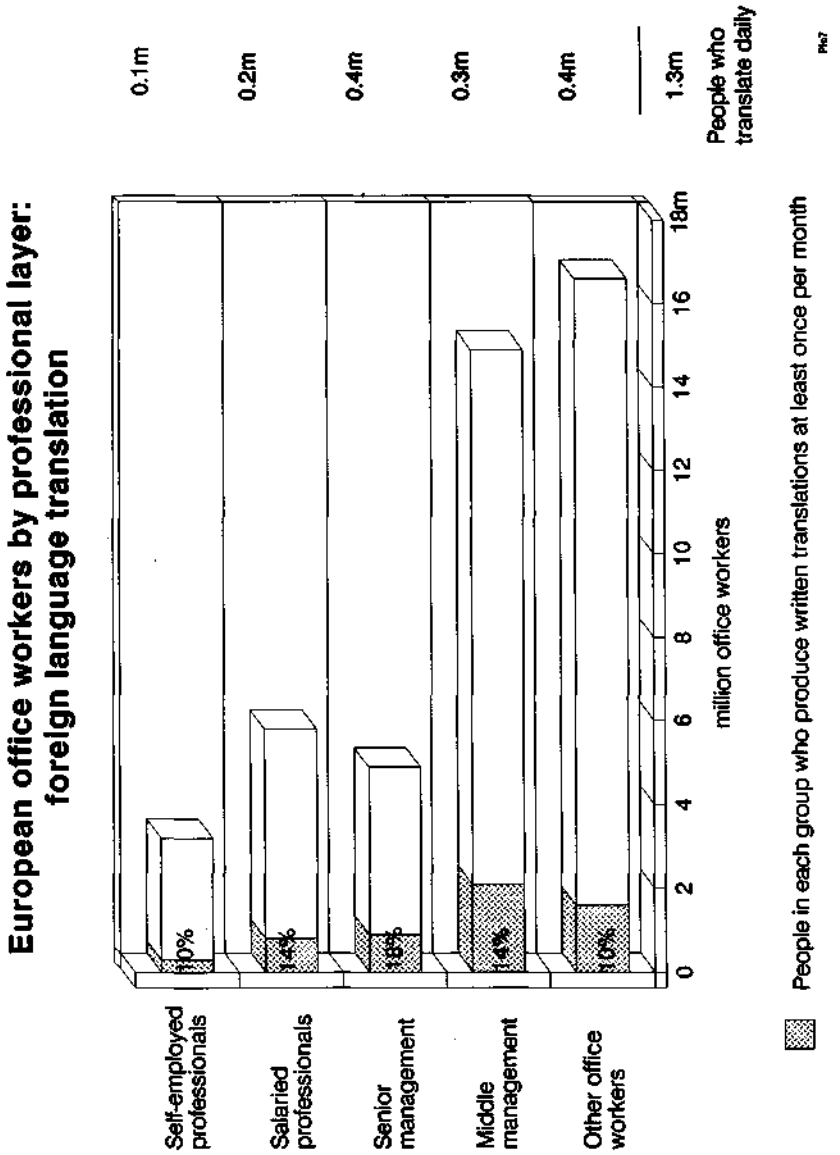

European office workers by professional layer:
foreign language translation
0.1m
0.2m
0.4m
0.3m
0.4m
1.3m
People who translate daily
10%
14%
18%
14%
10%
0
2
4
6
8
10
12
14
16
18m
million office workers
Self-employed professionals
Salaried professionals
Senior management
Middle management
Other office workers
People in each group who produce written translations at least once per month

Figure 3

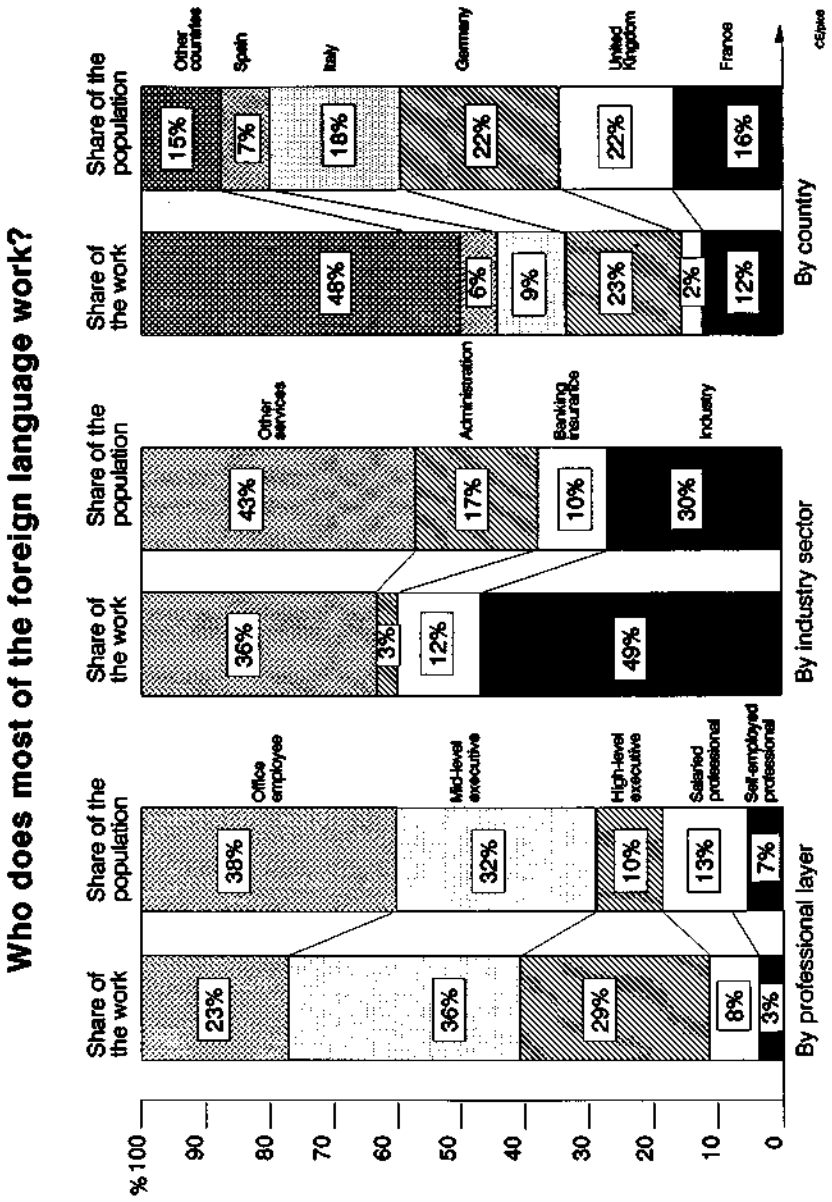

Who does most of the foreign language work?
Share of the work
Share of the population
By professional layer
Office employee
Mid-level executive
High-level executive
Salaried professional
Self-employed professional
23%
36%
29%
8%
3%
38%
32%
10%
13%
7%
By industry sector
Other services
Administration
Banking, insurance
Industry
36%
3%
12%
49%
43%
17%
10%
30%
By country
Other countries
Spain
Italy
Germany
United Kingdom
France
48%
6%
9%
23%
2%
12%
15%
7%
18%
22%
22%
16%
% 100
90
80
70
60
50
40
30
20
10
0

Figure 4

industry is using foreign languages most; administration least.

On the right is the same breakdown by country. As you see, not much foreign language activity happens in Britain! Other countries – including Denmark, Netherlands, Greece and Portugal who have the less common languages – do most work in foreign languages in relation to their size.

I don't know if any of these findings surprise you. When we presented them to our EC customer, he told us to make sure we had a plausible explanation ready for anything that looked in the least surprising:

> 'My colleagues at the EC', he said, 'when they look at survey results, always react in one of only two ways: if the numbers show what they expected, they say: "that's obvious"; if they don't, they say: "that's nonsense".'

LANGUAGE TECHNOLOGY PRODUCTS

Going on to the promised overview of language technology products, I have divided them into five groups:

- writing and editing aids
- translation tools
- natural language database interfaces
- text scanning systems, and
- talkwriters, or machine dictation systems.

This is really rather a simplification and does not do justice to the many different application areas which have been successfully tackled with language technology solutions.

This is not to say that the products currently on the market are all very sophisticated; some barely qualify as language technology tools at all; nor does it mean that there are many fully-fledged commercial products out there that you can buy off the shelf – there are actually very few.

LT products go through the same cycle as other software products:

- A new product starts with a one-off solution for a particular customer, which is usually very difficult to produce, and therefore very expensive.
- Next, suppliers solve similar problem situations for other customers; gradually, there are components that can be re-used, a product shell or a development workbench might exist, and solutions become cheaper. Nevertheless, each new system still needs intensive customisation from the software developer.
- Gradually, the software developer realises that this is no way to make money – he needs to turn the customisation process over to the customer. So he builds a user-friendly system around the

customisation process and tries to sell the whole thing as a product or package. At this stage, the developer must still rely on the goodwill of the user – if the user is not prepared to make the effort, there is no chance that the system will succeed.

— Finally, the product has been adapted to so many different situations that it becomes truly general-purpose. This is when it might become available in 'shrink-wrapped' form and sell in large numbers.

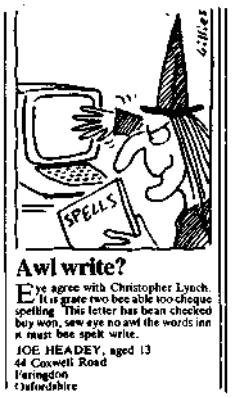
Unfortunately, because of the very difficult process involved in language technology, this progression has taken much longer for LT products than, for example, for database systems or word processing systems. Many of the products which are now on the brink of success are based not on years but decades of research. Acceptance has been slow; and there are as yet very few off-the-shelf products.

In talking about each of these product groups, I will give you a few examples and tell you how I think these might develop over the next five to ten years. And I will tell you, where this makes sense, how many potential users of these systems there are in the 12 EC countries according to the survey.

Writing and editing aids

This is the less sophisticated end of LT product scale, at least in its current state of development. In fact, it is difficult to decide sometimes what constitutes an LT product. The criterion for counting something as an LT product was that it should have – and use – some linguistic knowledge other than simple list matching (like most spell checkers). So – what products are there in this group?

Intelligent spell checkers – those which take some account of the context a word appears in. I am told there are some French products of this type, but have not been able to get any names.



Grammar and style checkers: the most successful appears to be Reference Software's Grammatik IV. There is now a British English version (as opposed to American), and French and German versions are in the pipeline. Grammatik checks grammar and style – on the style side, it checks for example for too many long words, for too much use of the passive form, and use of clichés.

MacProof is probably similar. There are French and German versions available, and even one that checks English written by French nationals – looking for typical mistakes a French person might make when writing in English.

Writing environments are integrated sets of programs that do all the above but also help the writer to structure his or her ideas. Cicile is one of these. It is not yet widely available.

There is a vast market for such products – there are 19 million Europeans who write and edit documents daily. This does not mean that I expect all these 19 million to be buying style checkers in the next year or two. First, these products must be widely available – in all the required languages – and at a very low price. Even then, our forecasts assumed only a 25 per cent penetration by the year 2000.

I expect that the simpler products will improve, and will gradually be bundled with word processing packages (as spell checkers and thesauri have been). They will cost the consumer very little extra in the end. The more sophisticated products may have a life of their own. The big breakthrough in style checkers will be when systems tell you not only what is wrong, but what to say instead! Such a system is said to be under development at the IBM research centre in Germany.

Translation tools

Three separate components here:

— terminology management packages (Superlex, TermTracer, Profilex);
— interactive translation systems (e.g. TransActive) which save the translator much time and effort, but require much input and interaction from the translator;
— machine translation proper which will produce at least an approximation to the final version of the translated text. Examples are Logos, Metal.

I said earlier that most of the LT products in existence today still rely heavily on the goodwill amongst potential users. Although it is possible to derive substantial benefit from LT tools, this will happen only if the user

wants the product to succeed. If he is out to find faults in a system, there are always some to be found.

I think machine translation has suffered from this more than most, and CAT and MT suppliers have had a very hard time gaining user acceptance for their systems.

Alpnet solved this problem by creating its own users – by buying up a large number of translation bureaux who now use Alpnet software for a substantial proportion of their translation work. The hope is that their larger customers will one day see the benefit of having their own in-house systems.

Logos, who seemed to be in difficulties recently, seems to have bounced back, having sold ten systems to the government of Ottawa. The company is now hoping to set up a network translation service – both in Germany, with the Deutsche Bank, one of its investors, and in the US, where AT&T will be the service provider. The service offered is billed as NEAT – N E A T – for Non-Edited Automatic Translation, and will be available over the telephone, with an overnight service. The theory is, and I can easily believe it, that there is a vast hidden demand for rough translation – good enough to give the reader an approximate idea of what a document is about – especially if it is cheap and fast. Again, if this catches on, it may encourage the purchase of more in-house machine translation systems ultimately.

I am told that there is such a service already on the French Minitel system – using Systran. I would be interested to hear of anyone who has experience of this!

There are theoretically 1.3 million potential customers in Europe for translation tools in general. These are not all professional translators, of course, and it is unlikely that a very large percentage of them will be taking to this type of translation tool.

On the other hand, this number ignores potential users of unedited translation over the telephone.

Natural language database interfaces

There are two main types of natural language interfaces to databases: those to numerical and those to text databases.

NL interfaces to numerical databases have been around for some time. Intellect and Parlance are examples, although they exist only in English. They are used in large companies which have masses of data available about their operations, but not in a form that is easy to get at by non-technical users. In the past, users who wanted a particular analysis or report had to ask their information services department to do it for them – and when the report arrived, often weeks later, it was either no longer relevant, or the user then realised that it wasn't really what he wanted.

Natural language interfaces

• Interfaces to numerical databases

Parlance	(BBN)
Intellect	(AI Corporation)
Q&A	(Symantec)

•Text database interfaces

Tome Searcher	(Tome Associates)
Realist	(Siemens)

Figure 5

Natural language interfaces give the user immediate access to management data. They allow users to type in perfectly ordinary questions, for example:

which department has achieved the highest sales this month?

The system will translate this into the SQL commands necessary to find all the data needed, to analyse it and to give the answer to the question.

There are also less sophisticated, PC-based NL interfaces to flat file databases. Symantec's Q&A is one such system; it exists in the major European languages and has been very successful.

Text databases – i.e. collections of research reports or newspaper stories – usually require users to be familiar with Boolean logic to retrieve the information they want. Tome Searcher and Realist are systems which help the users to formulate their enquiry. They are not strictly natural language interfaces, but go some way towards them.

All larger companies have both numerical databases (personnel, sales, invoicing) and text databases (correspondence). It is conceivable that in ten years' time these interfaces will have become so generalised that even low-level office employees could use them to do very mundane tasks, for

example, to retrieve a letter from the file, or to find out a customer address. If this happens, the potential market is very large indeed.

Text scanning systems

The basic idea here is that the system scans a body of text and understands enough about it in order to take appropriate action. This action might be to reformat the document, to send it to a particular recipient, or simply to file it in an appropriate place.

Atrans has been around for a while – it is used in banks and reads unformatted money transfer telex messages. It extracts the key information such as amounts, sender and recipient; it verifies bank and account reference numbers; and reformats the message in such a way that it can be processed easily, either manually or even automatically.

Obviously, this particular system can only work in this closely defined application area where the message content is quite predictable.

Another application is text indexing. TIS is a system developed by Carnegie Group in the US for Reuters in London, who have a widely-used text database of financial news stories. TIS analyses the contents of articles from 1,000 different publications and classifies them into several hundred topic categories, such as asset transfers, privatisations, or currency stories. It also indexes them by country and company name. This job used to be done manually. The new system produces classifications more quickly and more accurately.

This system was developed specifically for Reuters, but Carnegie Group is using the underlying technology – which it calls a text categorisation shell – to help other companies with their text classification requirements.

A number of other companies are working on intelligent indexing systems. The next logical step is the automatic production of precis or summaries of articles. It is conceivable that in ten years' time we will all have our incoming information sorted, summarised and perhaps even weeded out for us.

Talkwriters

• Free text dictation systems

DragonDictate3OK (Dragon Systems)

• Report writers

VoiceMed (Kurzweil Inc)

Total potential European customers: 16 million

Figure 6

Talkwriters

Talkwriters are computer dictation systems – basically typewriters that you talk to. I won't say any more about this category at the moment as I want to come back to it later – except that the market for them is very large indeed!

Figure 7 gives some tentative market forecasts for these five product groups in the year 2000: in value, these products will add up to over half a billion ecus. Database interfaces and talkwriters seem to have the greatest potential for growth. Translation tools will earn some 38 million ecus from European customers by 2000.

The talkwriter – a typewriter that you can talk to – is every IT supplier's dream product – they all *know* it will be a success if it can only be made to work. Figure 8 shows a user sitting back comfortably – with a microphone – talking to his PC, and watching the text appear on the screen. Many companies are working on it – in fact, when Tim Johnson wrote Ovum's NL report in 1985 two companies at least thought they would have a

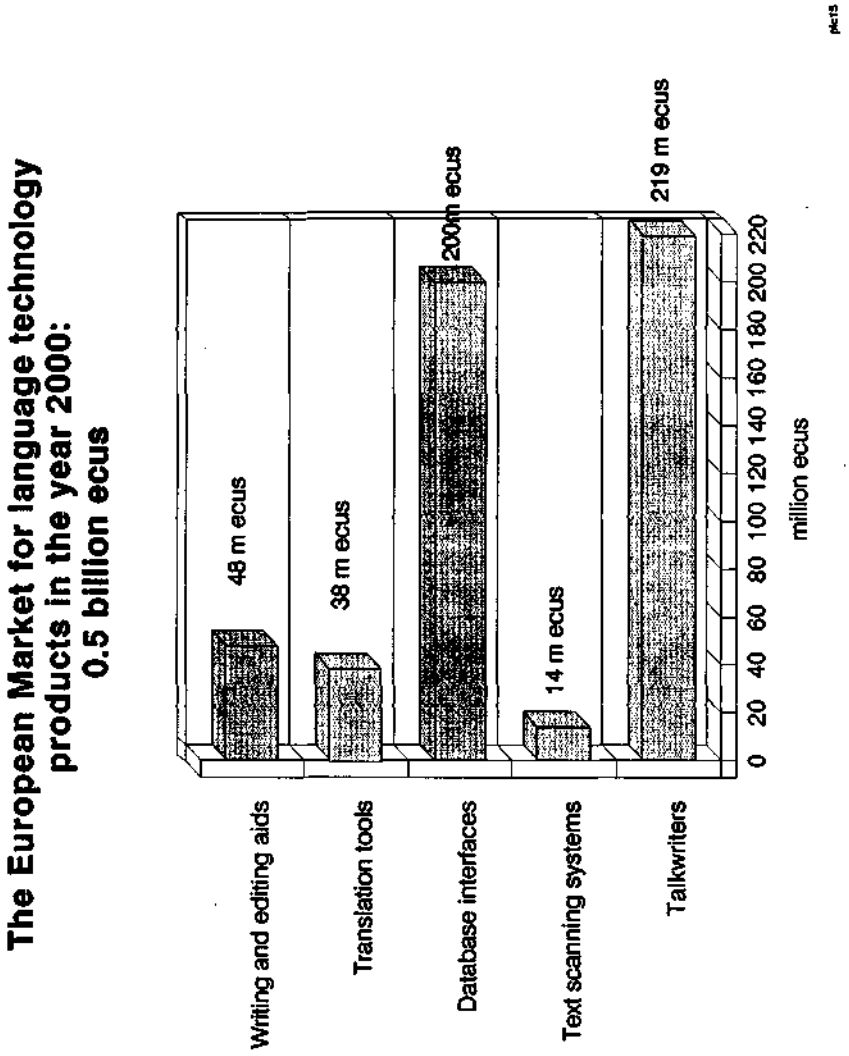

The European Market for language technology products in the year 2000: 0.5 billion ecus
48 m ecus
38 m ecus
200m ecus
14 m ecus
219 m ecus
Writing and editing aids
Translation tools
Database interfaces
Text scanning systems
Talkwriters
0 20 40 60 80 100 120 140 160 180 200 220
million ecus

Figure 7

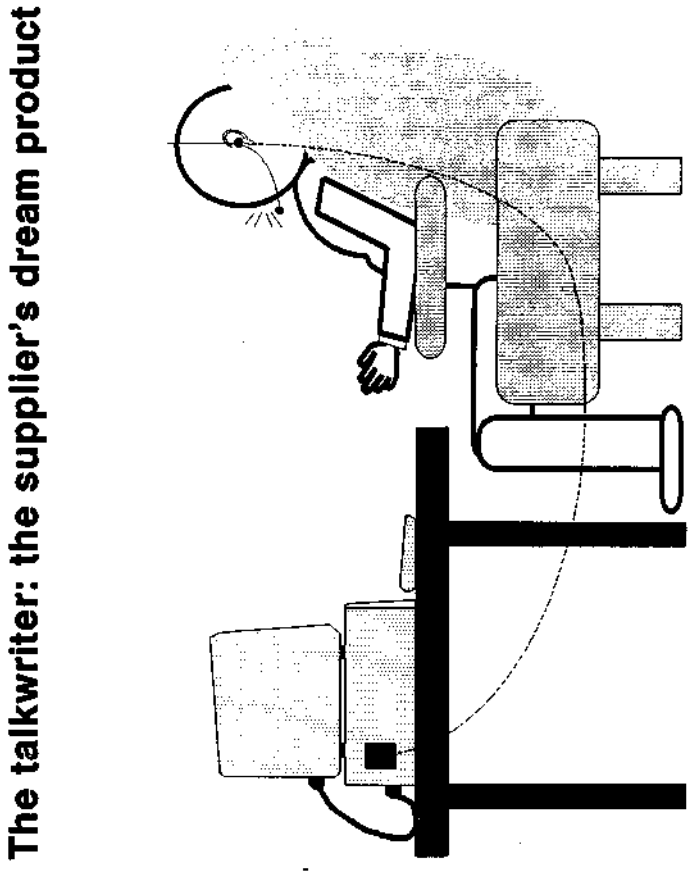


Figure 8

commercial product within a year; yet there is only one commercial product on the market today.

What are the problems?

There are five dimensions to talkwriter development: actually, if you look closely there are six. There is, first of all, the quality of voice recognition. But I put this in brackets as the actual signal processing is the part most suppliers have already solved satisfactorily.

Apart from voice recognition, there is, first of all, size of vocabulary. Obviously it is much easier to produce a system that understands only ten or 100 words rather than several thousand, but a very large vocabulary is needed for free text dictation.

Second, there is the matter of how many speakers a system will recognise. Again, one speaker – with *one* type of voice and *one* accent, is much easier to cope with than many different speakers.

Third, there is the question of continuous versus clipped speech. The job is made much easier if the system is told where one word ends and the next one starts; so some systems ask the speaker to make-tiny-pauses-between-words.

Fourth, there is accuracy. If the user has to spend more time editing the material afterwards than he did dictating it in the first place, he will soon lose interest.

Even with clipped speech, it is not enough to be able to recognise signals to identify the particular word the speaker wanted. As you know, there are many words that sound the same but are spelt differently. Also, as you all know from taking down names over the telephone, there *is* no acoustical difference between, for example, TALE and PALE or between SUN and FUN. You can tell them apart only by watching people's mouths, or from the context in which they occur.

So a talkwriter needs to know more than just the signal pattern of each spoken word to get it right. It needs to know not only what noises the user has made, but what he is likely to say next. However, taking the context into account is difficult for a talkwriter. Unlike machine translation, where the system can skip about in the sentence and gradually identify subject, object etc., the talkwriter has to make up its mind straight away as soon as the user has said another word: the user does not want to wait until he has spoken the whole sentence before seeing text on the screen. So a talkwriter can, at most, take into account the words that have gone before.

So you can see that producing a talkwriter is not a trivial problem.

Finally, there is price – no use producing a talkwriter that only works on an expensive supercomputer!

As I said, there are a number of companies devoting considerable research effort to this challenge. They all have talkwriter prototypes that solve part of the problem. It is generally estimated that it will take another five years before the real thing emerges – an affordable talkwriter that

understands any speaker, using natural continuous speech, without placing restrictions on the grammar or vocabulary.

Meanwhile, one small US company has decided to take on the market with an *Imperfect* system. This DragonDictate, which does have a very large vocabulary; it does understand any speaker; but you do have to use clipped speech, or discrete speech as they call it.

DragonDictate's accuracy is not perfect, either. Nevertheless, it works – because it has some very elegant ways of overcoming this problem!

As you dictate, each word you say appears on the screen. Quite often, DragonDictate is not sure what you have said. In that case, the word it thinks most likely will appear in the sentence on the screen. How does it know which word of all the possible words is the most likely? It does this by referring to its body of statistical knowledge – derived from analysing vast corpora of text – of which words frequently appear together.

In addition to the chosen word, there will be a little window in the top left-hand corner of the screen which gives you another ten words to choose from. If the word in the text is correct, the user just carries on. It *can* happen that the word DragonDictate chooses is nothing like the one the user said – but there, in the little extra list, perhaps six words down, *is* the one he wanted. He says: CHOOSESIX, and this word appears in the text, replacing the wrong one. If the word is not in the list, he can say: SCRATCHTHAT, and try again. If the system fails him again, he can say: SPELLMODE, and start spelling the word. Before he has got very far, there will be a new list of words to choose from: the system now knows not only what the word sounds like; it also knows the first letter or two, which is often enough to identify it. So the user rarely needs to spell the whole word.

DragonDictate runs on a PC; the software costs $9,000. People who use it can achieve speeds of between 20 and 45 words per minute. This in not fast enough to beat any proper typist, but it is miles better than typing with two fingers.

The first few systems were sold to handicapped people, to whom it is a real blessing; but it is now beginning to catch on in general business applications. So far it only understands English, but a Belgian company is producing other language versions for Dragon Systems. It will be very interesting to see if it succeeds, and what the large suppliers will do about it!

AUTHOR

Brigitte Engelien, Principal Consultant, Ovum Ltd, 7 Rathbone Street, London W1P 1AF, UK.