PROVIDING TERMINOLOGICAL RESOURCES FOR VOCATIONAL TRAINING in LWUTLs: The VOCALL Project[1]

Andrew WAY

Dublin City University

1. Description of the Project

This paper details work done on the **VOCALL** (Vocationally-oriented Computer-Aided Language Learning) project, funded under the *Leonardo* Programme, for the period December 1995 to August 1997. The **VOCALL** project is co-ordinated by Dublin City University (*DCU*), Ireland, and its other constituent partners include *ILTEC*, Lisbon, Portugal; *ILSP*, Athens, Greece; *top Schulung*, Hamburg, Germany; and *FAS*, Dublin, Ireland.

The aim of the project is to build language learning tools for vocationally-oriented learners in the areas of computers, business administration, and electronics. The multimedia product, which will be identical in all languages of the project, will be marketed as a self-learning tool for FL learners, as well as L1-disadvantaged learners, in vocational and professional training in the areas mentioned.

The reasons why such a tool is sorely needed are many, and include:

- fostering methods of self-learning in the workplace
- furthering access to vocational training
- promoting equality of opportunity
- adapting to industrial changes

The **VOCALL** project resulted from a merging of two consortia—one containing *DCU*, *ILSP* and *ILTEC*, and the other *FAS* and *top Schulung*. A primary focus of the first of these groups, which remains to this day, was that of less widely used and taught languages (LWUTLs), namely Portuguese, Greek and Irish[2]. In these cases, their maintenance as first language (L1) in vocationally-oriented language learning (VOLL), especially for disadvantaged users, cannot be separated from their promotion as foreign languages (FLs) for immigrant populations and learners in other member states. Written language resources such as lexica and terminology banks (let alone spoken language resources) are not well developed as learner aids in the case of LWUTLs, and these concerns are addressed in

[1] This work has been made possible through EU-funding under the *Leonardo* programme.

[2] English was included qua lingua franca. The addition of German has helped provide a focal point for the testing phase of the project, as the principal language studied by *FAS* trainees in these areas is German.

VOCALL in the development of a standardised series of learner aids for the improvement of linguistic skills in vocational contexts.

Areas such as Computer Technology and Electronics are developing rapidly and new forms of linguistic training and qualifications are needed on account of borrowing from English. Vocational and professional training with a linguistic requirement, be it in L1 or FL, are increasingly in demand, and this initiative will strongly support this development.

With this in mind, it was foreseen to source, develop (where necessary) and make available in digital format (disk, CD-ROM), using widely available technology (e.g. MS-Access, MS-Windows), a multilingual glossary of technical terms in the areas mentioned above, for the languages of the partners, as part of a self-learning tool encompassing multimedia technology (sound, video, graphics, text).

As well as the envisaged product being used as a self-learning tool in VOLL centres, we anticipate it being of interest to peripheral regions (in which FL competence is limited), small businesses, and in distance-learning centres. It will be of benefit to learners in an adult or lifelong context with a disadvantage in previous L1 or FL education, young learners in first-chance vocational education, as well as immigrant groups in member states whose national languages are LWUTLs.

2. Terminological Issues

The terminology work underpins all other aspects of the proposed tool. The glossaries themselves act as mnemonic devices, include fields for pronunciation of the words, contain hyperlinks to textual components based on the core language learning material, and are the basis for the multiple-choice self-testing tool built in conjunction with the video component.

Given this, it was imperative that we develop a methodology for performing this basic research which could be replicated for subsequent sublanguage areas. Furthermore, the lists had to be constructed in close consultation with language teachers as well as technical experts, both of whom also took part in validating the lists of terms.

2.1. Sourcing of Terms

It was agreed that the field of Computer Skills would be used for the pilot study, as it was felt that of the areas to be tackled, this was the most likely to have available terminology in all of the languages of the project. Two of our partners are training institutions: *FAS* in Ireland and *top Schulung* in Germany. It was decided that the corpus to be used for the pilot study would be the material prepared for the English computing courses taught at *FAS*. Given that we wanted to ensure that the tool was entirely relevant to *FAS* trainees, we were careful to bear in mind the possible end-user profiles, these being:

- School-leaver with junior-certificate qualifications only.
- School-leaver with leaving-certificate qualifications.
- Person who is re-training, having worked previously in another area.
- Person with degree picking up new skills.

As can be seen, there is a fair degree of disparity between these end-users. Given this, it was all the more imperative that the terminology to be integrated into the tool be corpus-based, so that we could be certain that all students would have had exposure to at least this core foreign language material contained in the *FAS* teaching materials. Without this constraint, attempting to pitch the tool at an appropriate level would have been an almost impossible task.

Firstly a corpus of texts was compiled in the area of Computer Skills, consisting of teaching materials supplied by *FAS*, for English. A list of terms was extracted from this corpus, and this list was sent to the other partners in May 1996. Further teaching materials were supplied by *FAS* in August 1996. The first list was then revised and a second, more complete list was sent to *ILSP*, *ILTEC* and *top Schulung* at that time, for consideration as to their suitability for inclusion in any final list. We shall describe this process below.

The same was attempted for the other languages. The German partner (*top Schulung*) researched the teaching materials of their various training centres, using the English list provided by *DCU* as a basic source of reference. They also used the English programs of *Word* and *Excel*, and a number of dictionaries, including one on-line, as reference materials. Our Portuguese partners (*ILTEC*) consulted vocational training colleges in Portugal, some of which provided them with their training material. Their corpus also consisted of handbooks on secretarial and office skills, as well as some terminologies.

For Greek and Irish, meanwhile, this process was rather more problematic. Our Greek partner (*ILSP*) encountered difficulties initially in sourcing teaching materials in Greek in order to build up a Greek corpus, as the teaching materials used by vocational training colleges in Greece are owned by the teachers themselves and not by the institutes in question, and the teachers were understandably reluctant to give us their notes (although this hurdle has been overcome to a certain extent in recent times). Given this, therefore, *ILSP* compiled a complementary list of English terms, which they sourced from two introductory manuals and from four computer lexica---three printed and one computerised. They also consulted the Help menus of the most commonly used software packages, e.g. *Windows 95, Word 7.0, Excel for Windows, Access for Windows*. Subsequently *ILSP* made contact with the *OAED* (Organisation of Manpower Employment) in Greece, which after the signing of a co-operation agreement between the two sites, enabled *ILSP* to be supplied with their teaching material, for the 2nd and subsequent terminologies.

For Irish, it was decided that even in the case of the computer terms, let alone the other areas addressed by the project, it would not be possible to build up a corpus given the lack of material available. Equivalent terms in Irish exist for approximately 50% of the terms included in the *DCU* English term list. New terms are being created by *DCU* for the remaining 50%. The newly created terms are submitted to *An Coiste Téarmaíochta* (The Terminology Committee for the Irish language) for approval and standardisation, this being the accepted practice for Irish term creation (see Uí Bhraonáin & Ní Dhubhghaill (1997), this volume).

In sum, where possible all language material was sourced from real teaching corpora. Where this was impossible, principled decisions as to what material should be used to obtain relevant terminology were taken in the light of the anticipated end-users, so that printed and computerised lexica, Help menus from well-known software packages, and advice from outside experts contributed usefully to this task. All partners used language trainers and other experts to validate the terminology produced. We also hope that a tentative agreement with Infoterm will lead to their involvement in the validation process for our databases.

2.2. Criteria for inclusion of terms in list/size of list

Once the initial lists had been gathered from the corpora available, we had to decide which terms would merit inclusion in the final, agreed list. All lists of terms created by the partners were collated into a merged list. The terms contained in the merged list were then tagged automatically so that a reference as to their source list remained (either individual partners' lists, or some combination thereof, as some words were included in more than one list). Some examples include the following:

abort	ILSPILTEC
active_window	DCUILSP
add_on_facility	DCU
backup	ALL

where "add_on_facility DCU" indicates that the term originated in the *DCU* list, "abort ILSPILTEC" shows that both *ILSP* and *ILTEC* lists contained the word "abort", and "backup ALL" means that "backup" was included in all three lists[3].

This merged list was then sent to all partners. Based upon all of the above material (*FAS* corpus, merged list and end-user profile information provided by *FAS*), each partner was able to specify a number of job-descriptions, on the basis of which the lists could be coded.

2.3. Coding of Lists

2.3.1 Computing Terms

Following a discussion of the end-user profile by electronic means, the following categories of words used in potential job descriptions for these users were decided upon:

A: secretarial /administration, data entry, telesales

B: systems maintenance

D: all basic computer terms

G: all general words

Some examples of these include:

cell	DCU	A
chart_wizard_box	DCU	A
'EPROM'	ILTEC	B
batch_processing	ILTEC	B
console	ILTEC	AB
character	DCU	D
'RAM'	DCU	D
analysis	DCU	G
authentication	ILTEC	G

As can be seen, combinations of the codes were deemed permissible where appropriate, e.g. "console ILTEC AB", where it was felt that the term "console", originally on the *ILTEC* list, belonged to both categories A and B.

---

[3] For this first terminology, the German list was invoked at a later stage, so was omitted from this initial process.

It was decided to adopt the strategy of distinguishing between terms per se, and general language vocabulary to be used with such terms. The first group (A-D) were subject to the coding methodology as outlined, while the general language terms (G) were extracted as simple, uncoded lists, with equivalent translations added.

The merged list was coded by different members of each site to ensure intra-site agreement. The resulting coded lists were then evaluated automatically according to a basic metric, summarised as follows:

1. All general language terms (G) extracted as lists, with equivalent translations added.
2. Any word contained in >=2 lists, automatically included.
3. Any word coded >=2 categories (A-D), automatically included.

This resulted in the "final" list of 938 terms. However, there remained some problematic cases once this list had been checked manually, namely:

- *Potentially outdated terms*: file manager, Printer Setup …
- *Inclusion of new terms*: ID field, counter data type ...
- *Duplications*: bit-binary digit, email-electronic mail ...
- *Synonyms*: toggle-toggle button, peripheral-peripheral device ...

The first group merited some discussion given that the project has a completion date of Dec 1998, and some of these terms may well be redundant by then. The 2nd group were problematic for the reason that they were very new, and it was difficult to say at this stage whether they would establish themselves as mainstream terminology during the lifetime of the project. The 3rd group were easier to deal with, given that we were convinced that these were duplicates. Nevertheless, the alternative terms would be used by different groups of people, so which one would we include? The final group were more difficult to deal with, given that in some circumstances they could be seen to be synonyms in context, but in others they may well be different parts of speech, for instance, in which case they would perhaps both need to be included.

Most of these problems were resolved in the manual checking of the list. Most potentially outdated terms were kept, as it was not thought to be detrimental to users to have available terminology relevant to older operating systems. New terms were mostly deleted, until such time as their widespread use merited their inclusion. Synonyms were the easiest group to resolve, in that what was considered the more available term was maintained, with others deleted. Duplications were treated by including, where possible, both "duplicates" as different syntactic categories.

2.3.2 Office Skills Terms

At the same time, work continued on the Office Skills database. A similar strategy was followed for this terminology also, but the coding process decided upon here was much simpler. Words were coded as to whether they were Basic (B) or Non-basic terms (NB) for each language. Again, once this had been done, the final list was derived automatically, i.e. any word coded B which was included in >=2 lists was included in the final list. This turned out to be 1111 terms, for English.

Equivalents were then sought for these terms in the other languages. However, it was not until this stage of the process that further problems were identified. For instance, there had been some confusion about the meanings of some of the terms, given that some

partners were obviously dealing with highly specialised terminology in a foreign tongue. In addition, the coding procedure itself had unfortunately been misinterpreted. For example, some terms had been coded as basic or non-basic for a specific language, whilst other coders had interpreted this task as a more general one, i.e. basic or non-basic for **all** of the languages of the project. Consequently, a 3rd category, namely "language-specific", was added. These words were deemed important depending on the language being studied, i.e. it would be inappropriate for our trainees to travel to such countries without being equipped with such vocabulary. Despite this, however, only the basic terminology would be included in the prototype, with the language-specific terminology added at a later stage.

As can be seen, the list derived via the automatic evaluation of 1111 terms was likely to contain errors. Nor had we defined end-users for this list and, as a result, each group was coding the list with a different set of end-users in mind. However, given the general way in which these terms were being coded, this was not felt to be a problem. There were also cases where a term could take more than one meaning but we had not discussed and agreed (as a group) upon which one should be the preferred reading in our list.

Consequently, the automatically derived lists, both for Computing and for Office Skills, were subjected to a final manual evaluation, using project partners in conjunction with outside experts, which provided the final lists which were input into the tool.

2.4. Content of terminological databases

Owing to these and other issues, it was agreed that the terms themselves (not the general language vocabulary) needed to contain at least the following information:

- Term
- Part of Speech
- Gender
- Pronunciation (audio clip, not phonetic transcription)
- Term Number (to link translations)

The part of speech tag was entered primarily for use by the project terminologists, as certain words can obviously appear as different syntactic categories (e.g. "file") in one language, which would necessitate different translations in other languages. However, given that this information is available, we decided to add a field for part of speech in the database entries as an end-user aid. Recall again that, given their vastly differing profiles, this would be of use to only a subset of our end-users. Likewise gender will be of use to more informed end-users, and will mean very little to those with less advanced language capabilities. The term number is irrelevant to the end-user; it is there primarily to aid the programmer to link together translations, as well as terms with their respective audio clips.

This latter facility, as already stated, is considered vital if our tool is to be successful. Given that some of these users view their **native** language faculties, let alone those in the foreign language, as something other than a skill to be used constructively in the workplace, we considered it imperative to focus more on the spoken rather than the written word. In these particular circumstances, this meant that on certain occasions, terms were included which differed somewhat from the mainstream (e.g. "Tippex", not "Correction Fluid"). Furthermore, given the profiles of typical end-users, it is highly likely that certain terms will only ever be used orally, i.e. it is probable that they will never see such terms in their written form, although again this facility remains available to end-users, should they find it of use.

We previously gave some consideration to the inclusion of examples as well as definitions. Whilst the former may well still be included, using a concordance tool to obtain their use in context directly from the language corpora, inserting definitions has been rejected outright, at least within the scope of this particular project. Whilst we are of the opinion that lists of terms per se are of limited use to our potential users, so that context needs to be provided, we felt that providing definitions, and more importantly ensuring consistency and compatibility with their translations in all of the languages of the tool, was too onerous a task, particularly given the time and money available to us.

2.5. Summary of methodology

We have presented above the steps involved in bringing about a methodology which we follow on the **VOCALL** project for deriving lists of terminology. This can be summarised as follows:

1. Derive initial term list in English from relevant course material
2. Source corpora in other languages
3. Seek equivalents for English terms in other languages
4. Propose additions to this list on the basis of foreign language material
5. On the basis of end-user profiles, code amended list
6. Perform automatic evaluation
7. Perform final manual evaluation

3. Current status of the multimedia product

We are currently just over halfway through the planned lifetime of the project. The proposed final tool will have several incarnations before it appears in its final state. Given that the tool will initially be tested by FL speakers of English and German, the Beta version exists for just these two languages, for the first two terminologies. The current contents of the tool are, therefore, as follows:

- 2000+ termlists for Computing and Business Administration, for each of the two languages
- Audio clips for all of these terms, for each of the two languages
- Core language material in textual format, based on the language material for these terminologies, with hyperlinks to termlists, audio clips, and the language testing section of the video component
- Video clips (editable by language trainers), as part of a multiple-choice self-testing component
- Two language games to stimulate the end-user

A testing scenario has been drawn up by experts in *FAS* and *top Schulung*, our training institutions, which will be rigorously implemented. This culminates in an exchange programme of trainees early in 1998, where we obviously hope that our tool will show an improvement in the language capabilities of the said trainees. Given a recent development at *FAS*, namely an increase in class size for the areas covered in this project, we feel that our tool has taken on an increasing importance if the technical vocabulary required is to be mastered by such students.

The other languages—Greek, Portuguese and Irish—will be incorporated in the tool by the end of September, again for the first two terminologies only. At the same time, work will continue in sourcing corpora in all languages for the Electronics area.

4. Further Work

The renewal report for continued funding of the project is due at the end of September. Assuming this additional funding to be forthcoming, following the interpretation of the results from our testing procedure, for all languages, we will amend the tool where necessary to ensure its suitability for our end-users, incorporating by next March our 3rd and final terminology. The tool then receives a second round of thorough testing, following which the final tool will be prepared by June. The final 6 months of the project will be spent producing the accompanying documentation and on-line Help, as well as the required reports and evaluation for the Commission.

When it comes to the dissemination and transfer of the experiences and the products of our project, each of the partners is prominent in their respective countries, and given that (for the most part) these are relatively small, the task of dissemination is unlikely to be onerous. Furthermore, Government agencies are likely to be interested in having a role in production and dissemination. The cost of the products will be kept to a minimum. With this in mind, the marketing of the product will begin in earnest next year.

5. Final Observations

This paper has set out to describe work being done on the **VOCALL** project, whose principal aim is to develop a vocationally-oriented CALL tool. Its primary focus was on the methodology developed and followed in the project in sourcing, coding and validating terminology. The paper also described the multimedia tool being developed, together with further work to be performed.

Given the modular design of the tool, we envisage the product being easily extensible to other languages and sublanguage areas. Furthermore, it would be simple to add on more facilities to the tool itself, or alter it in different ways to make it useful to other end-users.

By providing a tool for use in the learning of LWUTLs (both by native speakers as well as foreign learners of these languages), we anticipate that the principal impact will be a raising of the status of LWUTLs in the vocational training of their linguistically disadvantaged citizens, whether in first-chance or adult and continuing education. This will impact on all the sectors chosen. We hope the provision of efficient, innovatory teaching aids will raise the quality of learning and teaching, thus leading to improvements in all such areas as well as wider mobility of European citizens, particularly in the vocational context.

Finally, an associated goal of the project is to transfer to all the members of the existing network existing local/national solutions to problems common to LWUTLs in the area of VOLL, resulting in technologically innovative products for these languages and an enhancement of transnational co-operation in this area. We hope that this paper shows that we have begun that process, although there remains much more to do.

REFERENCES

Uí Bhraonáin, D. & A. Ní Dhubhghaill (1997): “Term creation for Irish-medium third-level education and vocational training”, in Proceedings of the International Congress on Terminology, Basque Centre for Terminology and Lexicography, San Sebastian.

****

LABURPENA / RESUMEN / RÉSUMÉ / ABSTRACT

Lanbide-Heziketarako baliabide terminologikoak ematea gutxi erabiltzen eta irakasten diren hizkuntzetan: VOCALL proiektua

VOCALL proiektua Europako Batzordeak sortu du *Leonardo* programaren babesean. Proiektuaren helburua lanbide-heziketako ikasleentzako hizkuntza ikasteko baliabide batzuk sortzea da, informatika, bulego-lan eta elektronikaren esparruetan eta, batez ere, gutxi erabiltzen eta irakasten diren hizkuntzetara zuzenduta, kasu honetan, irlandera, portugesa eta grekoa.

Gutxi erabili eta irakasten diren hizkuntzen kasuan, lexiko-banku eta terminologi bankuen moduko idatzizko baliabideak ez dira behar bezala garatu ikasleentzako laguntza-tresna izan daitezkeen aldetik. Eta hori gogoan hartuta, eta multimedia bidezko CALL izeneko halako egitura baten barruan, esparru horietako bitako termino teknikoekin glosario eleanitzak egin ditugu, partaideen hizkuntzetan. Tresna hori, gaur egun oraindik prototipoa dena, proiektuko hizkuntza guztietarako berbera izango da, eta hizkuntza nork bere kontura ikasteko tresna gisa merkaturatuko dugu, bai atzerriko hizkuntza (2H) ikasi behar dutenentzat, bai lehen hizkuntza (1H) ikasteko arazoak dituztenentzat ere, lanbide-heziketan eta aipatutako esparruetan, beti ere.

Artikulu honetan, terminologia eleanitza sortzeko metodologia bat proposatuko dugu, orain arte izan ditugun esperientzietan oinarrituta; baina tresnaren beste alderdi batzuk ere azalduko ditugu.

Provisión de herramientas terminológicas para la formación profesional en lenguas minoritarias en cuanto al uso y a la enseñanza: proyecto VOCALL

El proyecto **VOCALL** es un proyecto fundado por la Comisión Europea en el marco del Programa *Leonardo*. El objetivo de nuestro proyecto es crear herramientas de aprendizaje de la lengua para estudiantes de formación profesional en las áreas de informática, secretariado y electrónica y está dirigido a lenguas minoritarias en cuanto al uso y a la enseñanza, en este caso, al irlandés, portugués y griego.

En el caso de este tipo de lenguas minoritarias, los recursos en lengua escrita como bancos léxicos y terminológicos no han sido suficientemente desarrollados como apoyo para el estudiante. El producto multimedia será idéntico para todas las lenguas del proyecto, y se comercializará como una herramienta de autoaprendizaje para estudiantes de lengua extranjera, así como para estudiantes no aventajados en la L1, dentro de la formación profesional y en las áreas arriba señaladas.

Nuestro artículo propone una metodología para la creación de una terminología plurilingüe tomando como base nuestras experiencias vividas hasta la fecha dentro del proyecto.

Mise à disposition de ressources terminologiques pour la formation professionnelle en LWUTL: le projet VOCALL

Le projet **VOCALL**, créé à l'instigation de la Commission européenne dans le cadre du programme *Leonardo*, cherche à élaborer des outils d'apprentissage pour des apprenants orientés vers des filières à formation professionnelle dans les domaines de l'informatique, des compétences de bureau et de l'électronique, en s'intéressant tout particulièrement des langues aux moins usitées et enseignées (LWUTL), en l'occurrence l'irlandais, le portugais et le grec.

Les ressources écrites langagières telles que les banques lexicales et terminologiques n'ont pas fait l'objet de développement véritablement satisfaisant comme aides aux apprenants dans le cas des LWUTL. Dans cet esprit, nous avons entrepris la compilation de glossaires multilingues de termes techniques dans deux des domaines cités (à ce jour) pour les langues des partenaires, en tant que partie d'un ensemble multimédia CALL. Cet outil, pour l'heure à l'état de prototype, qui sera identique pour toutes les langues du projet, sera mis sur le marché comme outil d'autoapprentissage aussi bien à l'intention d'apprenants de langue étrangère (FL) que d'apprenants connaissant des difficultés avec leur première langue (L1), en formation professionnelle et occupationnelle dans les domaines mentionnés.

Cet article proposera donc une méthodologie pour la recherche de terminologie multilingue basée sur notre expérience, à ce jour, tout en rapportant sur d'autres aspects de l'outil.

Providing terminological resources for vocational training in LWUTLs: the VOCALL project

The **VOCALL** project, funded by the European Commission under the *Leonardo* Programme, seeks to build language learning tools for vocationally-oriented learners in the areas of computers, office skills and electronics, and focuses particularly on less widely used and taught languages (LWUTLs), in this case Irish, Portuguese and Greek.

Written language resources such as lexica and terminology banks are not well developed as learner aids in the case of LWUTLs. With this in mind we have compiled multilingual glossaries of technical terms in two of the given areas (so far) for the languages of the partners, as part of a multimedia CALL package. This tool, currently at the prototype stage, which will be identical for all languages of the project, will be marketed as a self-learning tool for foreign-language (FL) learners, as well as disadvantaged learners of their first language (L1), in vocational and professional training in the areas mentioned.

This paper will propose a methodology for sourcing multilingual terminology based upon our experiences to date, as well as reporting on other aspects of the tool.