Using Richly Annotated Trilingual Language Resources for Acquiring Reading Skills in a Foreign Language

Dragoş Ciobanu, Tony Hartley and Serge Sharoff

Leeds University Centre for Translation Studies
Leeds, UK
E-mail: smldc@leeds.ac.uk, a.hartley@leeds.ac.uk, s.sharoff@leeds.ac.uk

Abstract

In an age when demand for innovative and motivating language teaching methodologies is at a very high level, TREAT - the Trilingual REAding Tutor - combines the most advanced natural language processing (NLP) techniques with the latest second and third language acquisition (SLA/TLA), as well as computer-assisted language learning (CALL) research in an intuitive and user-friendly environment that has been proven to help adult learners (native speakers of L1) acquire reading skills in an unknown L3 which is related to (cognate with) an L2 they know to some extent. This corpus-based methodology relies on existing linguistic resources, as well as materials that are easy to assemble, and can be adapted to support other pairs of L2-L3 related languages, as well. A small evaluation study conducted at the Leeds University Centre for Translation Studies indicates that, when using TREAT, learners feel more motivated to study an unknown L3, acquire significant linguistic knowledge of both the L3 and L2 rapidly, and increase their performance when translating from L3 into L1.

1. Introduction

The need to devise new, more effective and motivating methodologies for language teaching and learning has become a priority of language tutors, researchers and decision makers alike. Demand exceeds the supply of such courses and adult education in particular requires serious attention (Chisholm et al., 2004; Colpaert, 2004). Technology also needs to be applied more carefully to the needs of language teaching, because it seems that in the last 25 years computers have not been put to good use in order to support language progress (Barrière & Duquette, 2002; Plass et al., 2003; Rouse & Krueger, 2004).

Moreover, a lot of attention is paid to phonetic and grammar exercises at the expense of acquiring reading skills, despite evidence from research that this skill is valued more than any other: "the majority of 167 distance teaching organisations [...] regarded reading and understanding the foreign language as the most important study aim" (Holmberg, 2005). For reasons such as lack of resources or inadequate training of tutors and learners in the use of ICT, it is often the case that language curricula do not provide enough time for the development of reading skills (Hunt & Beglar, 2005), although reading has also been proven to benefit many other areas of language learning (Pressley in Grabe & Stoller, 2002; Sun, 2003).

This paper demonstrates how a novel language learning methodology can be designed and implemented within a web-based environment using both new and already available linguistic resources. Specifically, our system is designed to help adult speakers (language L1, here English) acquire reading skills in a foreign language (L3, here Romanian) that is cognate with a second language they know to some extent (L2, here French). TREAT (Trilingual REAding Tutor) dynamically processes user requests to provide learners with linguistic information extracted from the corpora that is intended to facilitate reading comprehension.

2. Research Questions

Our first hypothesis was that a multilingual, corpus-based reading model that provides users with extensive reading materials and other relevant linguistic information extracted using natural language processing (NLP) techniques is more effective than traditional instruction in helping users acquire reading skills in an unknown L3 which is typologically related to an L2 they have some knowledge of.

Secondly, we hypothesised that given an effective learning environment, users can acquire the lexical and grammatical features of the target L3 without explicit instruction.

Thirdly, we aimed to show that reading resources can be arranged automatically in multilingual clusters that can boost the user's background knowledge to the level necessary for completing reading tasks successfully. To our knowledge, TREAT is the first computer-assisted language learning (CALL) application to demonstrate this.

The fourth hypothesis was that, by involving the L2 in the process, learners will both perceive and appreciate its support function, and seize the opportunity to use and improve their L2.

3. Methodology

Although there have been few attempts to create a framework for teaching and learning L3s that belong to the same language family as an L2 learners are somewhat familiar with - e.g. the EuroCom*Rom* initiative (Klein et al., 2002), whose main am was to show users "How to read all the Romance languages right away"-, the deliverables of such projects have fallen short of expectations and the evidence on which they were based was often anecdotal.

We have built on the latest research in the fields of second and third language acquisition (SLA/TLA), as well as NLP and CALL. Thus, we have bridged the frequently-mentioned gap between teaching practitioners, researchers and computer specialists (Felix, 1997; Barrière & Duquette, 2002; Borin, 2002; White, 2005; Yeh & Lo, 2005).

Our approach allows users to acquire background knowledge by exposing them to multilingual related reading materials, enabling them to select reading materials according to their individual interests, as well as formulate and validate linguistic hypotheses using evidence extracted from authentic corpora, and acquire knowledge about the L3 - as well as the L2 - vocabulary and grammar.

Moreover, we have also taken our methodology outside the laboratory and into the real language teaching and learning world – which is one of Chapelle's (2004) main recommendations - by implementing it in a web-based environment – TREAT -, which we then tested with the help of MA students in Applied Translation Studies. The results show that our approach is both more motivating and more effective than traditional language learning methodologies.

4. Resource Creation and Processing

4.1. Corpus Creation and Processing

We have assembled ad-hoc, comparable corpora of on-line news items in English (131 articles), French (100) and Romanian (182). In terms of size, we have been working with 81,812 L1, 85,342 L2 and 71,199 L3 tokens.

The original HTML files were automatically processed to discard boilerplate text and preserve only the news article. The results were POS tagged and lemmatised using TreeTagger for the L1 and L2 corpora. In the case of the L3 corpus, the Romanian Academy Centre for Artificial Intelligence (RACAI) had developed a language model for the TNT tagger (Tufiş, 2000); further work was put into improving the latest language model in order to use it for tagging and lemmatising the L3 articles.

We also used the English and Romanian WordNets together with a list of 1,766 English - French cognates[1], which altogether provided L3 synonyms, L1/L2 equivalents, L1/L2/L3 related words and L1/L3 definitions for 62% of the noun, adjective, verb and adverb lemmas in our L3 corpus.

In order to increase the support for our users, and given that our environment offers them sufficient authentic materials to verify the validity of any inferences, we also used a freely available string-similarity Perl module[2] in order to identify which L1 and L2 corpus tokens and lemmas are similar to L3 lemmas. This way, our environment provides assistance in the case of a further 29% of all content lemmas – i.e. nouns, adjectives, verbs and adverbs. A qualitative study performed on a random sample of 10% of these 29% L3 lemmas that were not initially covered by the L1 and L3 WordNets revealed that, in 62% of cases, the set of L1 and L2 structurally similar words and lemmas that were automatically identified contained sufficient cognates of the L3 target lemma.

Therefore, by adding this last resource, we have succeeded in providing our users with helpful linguistic information in the case of over 80% of L3 content lemmas.

[1] http://french.about.com/library/vocab/bl-vraisamis-a.htm

[2] http://search.cpan.org/~mlehmann/String-Similarity-1.02/Similarity.pm

4.2. Automatic Related Article Identification

Another prominent aspect in the multilingual corpora processing stage is the identification of related articles in L3, L2 and L1. The first phase consisted of computing relative frequencies for all L1 / L2 / L3 lemmas both in the L1 / L2 / L3 corpora, respectively, and in each L1 / L2 / L3 article. We were thus able to identify important lemmas for each article based on an empirically-tested threshold – i.e. if the lemma was a content one and if its relative frequency was 5 times greater within an article than within the particular language corpus, it would be judged as important for that text.

During the second phase, we used the L1 and L3 WordNets together with the list of L1-L2 cognates in order to compile three lists of important lemmas – in L1, L2 and L3 respectively - for each L3 article. In the case of the L3 list, it was made up of important lemmas together with their synonyms as suggested by the L3 WordNet. For L1 and L2, the lists consisted of equivalents of the important L3 lemmas.

The third phase was represented by the identification of important L1 and L2 lemmas for each L1 and L2 article respectively, by comparing their relative frequency in the article with that in the entire L1/L2 corpus.

Finally, we performed an intersection of these important lemma lists and set an empirically-tested threshold in order to identify suggested related articles (SRAs) in all three languages of the project. The formula we used was: $2xy/(x+y)>=T$, where xy represents the number of common important lemmas between articles 1 and 2, $(x+y)$ is the sum of the number of important lemmas in the two articles, and T is our threshold. Then we sorted the related articles starting with the one with the highest score. Figure 1 is a graphical representation of the related article identification process in L3. In this example, only L3 articles 2 and 4 are identified as being related, since only the results of the formulas $2*A1A2/(A1+A2)$ and $2*A1A4/(A1+A4)$ are greater than or equal to the minimum threshold T.

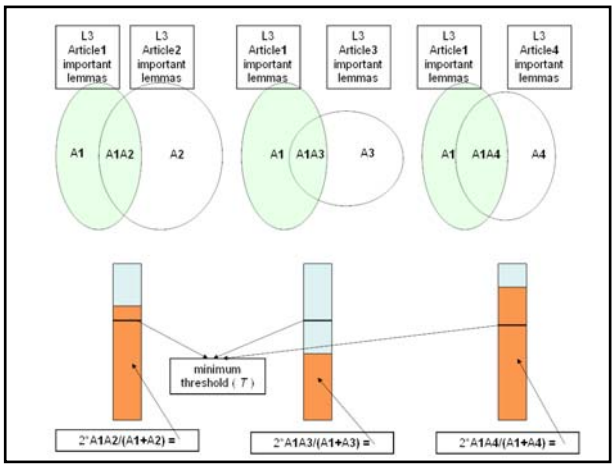

Figure 1: Automatic identification of L3 related articles

These resources allow users engaged in reading an L3 article to refer at any time to articles on the same/similar topics in L1, L2 and L3 in order to build up their background knowledge and notice multiple instances of authentic usage of familiar/unfamiliar vocabulary in all the project languages. To this end, we have also built a

customised concordance tool, which we present in section 5.

We performed a qualitative evaluation of the accuracy with which our tool identified authentic related articles (ARAs) for each L3 text in our multilingual, comparable corpus. We analysed a random sample of 50 L3 articles out of the total of 182. We looked at the top 5 SRAs for L3, L2 and L1 and noted the position in which the first ARA was, as well as the percentage represented by ARAs out of the first 5 suggested. If the first ARA was in the first position, we gave it 5 points; if it was third, we gave it 3 points; if none of the 5 SRAs turned out to be ARAs, the score was 0. Table 1 presents our results: *S* represents the average score for the first ARA; *StDev* represents the standard deviation from this score; and *P* represents the percentage of ARAs among the first 5 SRAs.

	S	***StDev***	***P***
L3	4.4	1.439	70%
L2	3.9	1.723	70%
L1	4	1.774	52%

Table 1: Accuracy of automatic related article identification

These results show that, overall, users can easily find ARAs at the top of the list of SRAs, as well as the fact that, at almost any time, 2 or 3 of the top 5 SRAs will be ARAs. This gives learners easy access to several reading resources on the same/similar topic and enables them to become familiar with relevant target vocabulary and structures more quickly.

5. Features of Treat

5.1. Article Selection Criteria

In order to enable users to identify the most pertinent resources, we have built several article selection criteria into our web-based interface. Initially we tested the most popular mechanisms for assessing the readability of a particular piece of writing – i.e. the Kincaid formula, the Flesch reading ease formula, and the Fog Index. However, such formulae rely on the length of words and sentences and, as our experiments have shown, our users had very few problems understanding and translating long L3 words and sentences. Therefore, we judged these formulae unsuitable for our purpose and made available a new set of criteria which include: article length; average sentence length; publication date; the occurrence of a particular part of speech in an article significantly more frequently than in the entire L3 corpus; the lexical density score; the number of SRAs in L3/L2/L1 or in all of these languages; the percentage of L3 content lemmas covered by the L1 and L3 WordNets together with the L1-L2 wordlist; and the domain.

Once the user selects an L3 article that suits their interests, the interface enables him/her to both read the article and browse through the suggested related articles in order to increase their background knowledge (Figure 2 demonstrates this very aspect).

Figure 2: Accessing SRAs in three languages from within TREAT

5.2. Concordance Window

In order to enable users to check their hypotheses about the way in which the target L3 works, we designed a custom multilingual concordance tool.

We are aware that the users' knowledge of the L3 is underdeveloped at this stage, yet SLA and TLA research indicates that even in the case of minimally useful contexts, even low-ability students can find useful information and can get an insight into the workings of the L2/L3, namely "word form, affixation, part of speech, collocations, referents and associations, grammatical patterning, as well as global associations with the topic" (Nation in Hunt & Beglar, 2005).

Our concordance engine enables users to perform searches in all three languages and, depending on the availability of WordNet information and current corpus data, it returns concordance lines in up to three languages, as well as relevant linguistic information about the target L3 word – i.e. POS, lemma, L3/L2/L1 synonyms /equivalents and related words, L2/L3 definition(s), L1/L2 structurally similar tokens, and L3 collocations. For each L1/L2 word search, the engine also consults L3 data in order to identify and return L3 vocabulary that is equivalent/related to the L1/L2 target one. In order to increase the performance of the engine in this case, we use L1 and L2 lemma information. Figure 3 presents the results of a search for the L3 word *uragan.*

Figure 3: Concordance window in TREAT

6. Testing and Evaluation Stage

We performed two types of evaluations: first of all of the users' performance and secondly of their experience with TREAT.

For the first one, we conducted a small-scale test by asking two groups of student volunteers to carry out a series of tasks. The first group (G1) was made up of eight individuals: seven MA students in Applied Translation Studies - some of whom were already familiar with L3 either through their own previous attempts to learn it or through having spent short periods of time in Romania -, together with one professional translator who had agreed to take part in our experiment. The second group (G2) also consisted of eight members, all MA students in Applied Translation Studies who were completely unfamiliar with L3

Both groups were asked to perform the same translation task T1, which involved rendering the same L3 segments into L1. The only difference was that G1 had to rely on their own knowledge, as well as Internet resources such as glossaries or bilingual dictionaries which more often than not did not support the use of Romanian diacritics. On the other hand, G2 had access to TREAT.

T1 was part of a larger set of tasks which included reading an L3 article and understanding it with the help of the automatically identified related articles, as well as our custom query engine. Furthermore, the users were asked to translate a number of L1 lexical items into L3 in order to give them the opportunity to use to a full extent all the functionalities of TREAT, and thus notice salient L3 morphological and grammatical elements. Finally, they were also required to scan, skim and summarise information.

The evaluation of the users' performance in the translation tasks was done by independent reviewers – native L1 speakers – who had access to the users' randomised and anonymised translations, and graded them for content (by comparison with a gloss provided by a native L3 speaker) and style (indicating how natural the translations sounded in L1). The maximum score awarded was 5, and the minimum was 1 - see Figure 4 for a comparison of G1 and G2's performances.

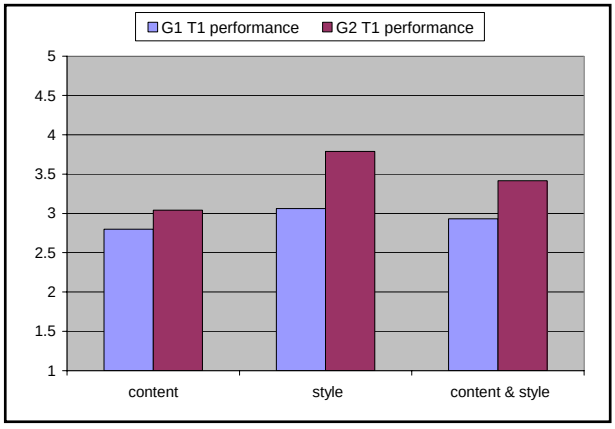

Figure 4: Comparative evaluation of G1 and G2 performance for task T1

Despite having less previous knowledge of L3 than G1, G2 scored higher both in terms of content and style, their performance being up to 25% better than that of the first group. Furthermore, the time factor was also an important feature, as in the professional world not only do translators need to produce quality work for clients, but they also need to do so as quickly as possible. Using TREAT, G2 members were able to finish their tasks in up to 50% less time than G1.

G1 were then introduced to TREAT, as well. This introduction consisted of an explanation of the architecture of TREAT, as well as a practical demonstration of its features, and lasted less than 30 minutes. G1 then proceeded to work on more tasks – some of them involving once more translations from L3 into L1, this time using TREAT. Figure 5 presents their progress.

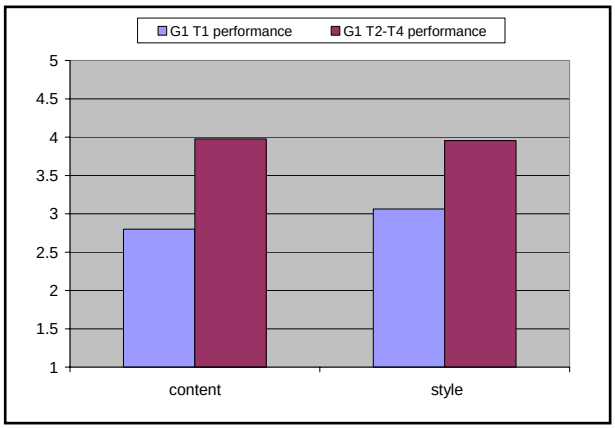

Figure 5: Comparative evaluation of G1 performance for task T1 vs T2-T4

The second type of evaluation was carried out by asking the members of G1, who overall had the longest experience using TREAT, to answer a questionnaire designed to elicit feedback on the perceived usefulness of our CALL environment functionalities, as well as verify our initial hypotheses about the effectiveness of our approach and its practical implementation.

The analysis of responses revealed that all the users:

- thought that the approach, as well as our learning environment TREAT, were original and motivating
- believed the approach was suitable to teach professional translators to read in a foreign language
- analysed thoroughly the L3 linguistic information provided by the TREAT concordance engine
- had used their L2 when working on our proposed
- would be willing to use this approach in conjunction with other course materials to learn other foreign languages

Moreover, the majority of users:

- would recommend this approach for learning to read in a foreign language
- believed the approach useful for university students and academics
- changed their initial attitude towards the L3 into a more favourable one

- thought that it is possible to improve three languages at the same time by using an environment such as TREAT
- found the interface and resources we had provided useful
- found TREAT very easy to use and easy to become familiar with (the rest found it relatively easy to use)
- found the L3, L2 an L1 SRAs useful and relatively useful, and consulted them occasionally
- found the concordance engine relatively useful (the rest of our users, however, stated that they found it useful)
- analysed the L1, L2 and L3 concordance lines very often
- found the suggested structurally similar L2 tokens relatively useful
- believed they had acquired knowledge of the L3 grammar and morphology
- thought they had improved their command of L2 to some extent

Finally, a minority of users (40%):

- used the article selection criteria we provided to find texts that suited their preferences better
- used the concordance engines to look up L1 or L2 words, too

Last, but not least, among the comments our users made, there was an indication that the accuracy of the automatic identification of structurally-similar L2 tokens needed to be improved, as it was occasionally misleading, especially in the case of function words, for which no coverage was provided by the WordNets. Moreover, we were also told that our article selection criteria appear sensible and useful, and that users are interested in working with them more closely in the future.

7. Conclusions

Despite the small size of our corpora, the results are promising. We have built TREAT in order to verify the validity of our research hypotheses. Having done that, the environment can now be expanded to work with larger corpora, or adapted to incorporate other language combinations. Any of the available corpus-gathering tools could be used to assemble more authentic linguistic data (Sharoff, 2006).

We have also proven that, even without heavy multimedia content, effective, scalable and easily maintainable learning environments can be built to motivate language learners to use their previous knowledge of L2's in order to learn completely new L3's. As far as usability was concerned, TREAT was judged as user-friendly and intuitive.

We are also encouraged by the fact that our users did indeed perceive an improvement in their command of the L2, which is solely a result of using TREAT and has not been reported in any other CALL or language learning study that we are aware of.

8. Bibliographical References

Barrière, C. & Duquette, L. (2002). Cognitive-Based Model for the Development of a Reading Tool in FSL. Computer Assisted Language Learning, 15(5), 469--481.

Borin, L. (2002). What have you done for me lately? The fickle alignment of NLP and CALL. In Proceedings of EuroCALL, Finland.

Chapelle, C. A. (2004). Technology and second language learning: expanding methods and agendas. System, 32(4), 593--601.

Chisholm, L., et al. (2004). Lifelong learning: citizens' views in close-up - Findings from a dedicated Eurobarometer survey. Luxembourg, CEDEFOP - The European Centre for the Development of Vocational Training.

Colpaert, J. (2004). Design of Online Interactive Language Courseware: Conceptualization, Specification and Prototyping. Research into the impact of linguistic-didactic functionality on software architecture. Doctoral thesis, Universiteit Antwerpen.

Felix, U. (1997). In the future now? Towards meaningful interaction in multimedia programs for language teaching. In F-J. Meissner (Ed.), Interaktiver Fremdsprachenunterricht, Wege zu authentischer Kommunikation (pp. 129--143). Tübingen: Gunter Narr Verlag.

Grabe, W. & F. L. Stoller (2002). Teaching and Researching Reading, Longman.

Holmberg, B. (2005). Teaching Foreign Language Skills by Distance Education Methods: Some Basic Considerations. In B. Holmberg, M. Shelley & C. White (Eds.) Distance Education and Languages. Multilingual Matters.

Hunt, A. & D. Beglar (2005). A framework for developing EFL reading vocabulary. Reading in a Foreign Language, 17(1).

Klein, G., et al., (Eds.) (2002). EuroComRom - The Seven Sieves: How to read all the Romance languages right away. Aachen, Editiones EuroCom.

Plass, J. L., et al. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. Computers in Human Behavior, 19, 221--243.

Rouse, C. E. & A. B. Krueger (2004). Putting computerized instruction to the test: a randomized evaluation of a scientifically based reading program. Economics of Education Review, 23(4), 323--338.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), WaCky! Working Papers on the Web as Corpus. Bologna, Gedit.

Sun, Y. C. (2003). Extensive reading online: an overview and evaluation. Journal of Computer Assisted Learning, (19), 438--446.

Tufiş, D. (2000). Using a Large Set of EAGLES-compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging. In Proceedings of the International Conference on Language Resources and Evaluation LREC 2000, Athens

White, C. (2005). Towards a Leaner-based Theory of Distance Language Learning: The concept of the Learner-Context Interface. In B. Holmberg, M. Shelley

& C. White (Eds.) Distance Education and Languages. Multilingual Matters.
Yeh, S.-W. & J.-J. Lo (2005). Assessing metacognitive knowledge in web-based CALL: a neural network approach. Computers & Education, 44(2), 97—11.