Corpus-based Tools for Computer-assisted Acquisition of Reading Abilities in Cognate Languages

Svitlana Kurella, Serge Sharoff, Anthony Hartley
Centre for Translation Studies, University of Leeds, UK
{s.kurella,s.sharoff,a.hartley}@leeds.ac.uk

Abstract

This paper presents an approach to computer-assisted teaching of reading abilities using corpus data. The approach is supported by a set of tools for automatically selecting and classifying texts retrieved from the Internet. The approach is based on a linguistic model of textual cohesion which describes relations between larger textual units that go beyond the sentence level. We show that textual connectors that link such textual units reliably predict different types of texts, such as 'information' and 'opinion': using only textual connectors as features, an SVM classifier achieves an F-score of between 0.85 and 0.93 for predicting these classes. The tools are used in our project on teaching reading skills in a cognate foreign language (L3) which is cognate to a known foreign language (L2).

1.Introduction

Recent developments in computer assisted language learning (CALL) focus on supporting the learning process through new media, e.g., using the Internet for developing communicative skills, or developing interactive multimedia materials for different goals, ranging from grammar training to complete language courses. The data driven learning (DDL) approach suggests supporting learning by using authentic digital resources related to the particular interests of learners, and by exploiting concordances and authoring tools in order to research and create language materials. However, while corpus linguistics generally is gaining in importance, research specifically into the use of comparable corpora for learning cognate (i.e. closely related) languages remains at an early stage.

This project aims at developing a methodology for English (L1) speakers to acquire reading competence in a third language (L3, here Ukrainian), based on some prior knowledge of a second, cognate language (L2, here Russian). The research is based on semi-automatic collection of corpora from the Internet, their annotation and the development of supportive learning methods. While others have primarily addressed lexical aspects of cognate L3 learning (e.g., Ciobanu et al., 2006), the focus here is on uncovering the semantico-logical text structure and genre-specific organization patterns, supported by formal text organization, that help promote successful reading strategies.

The goal is to develop an approach which can further be applied to any pair of Slavonic languages, and potentially to any pair of historically related and structurally similar languages within language families.

In this paper we present our approach to data-driven language learning and compare it to existing research (Section 2). Then we discuss the use of connectors in guiding the acquisition of reading strategies and, at the same time, in enabling automatic classification of texts (Section 3). In Section 4 we discuss further possible developments within the proposed framework.

2.Data driven language learning

It is well known that a foreign language learner benefits from previous experience of learning any other foreign language. It is assumed that L3 – more generally Ln – learners are more confident and successful on the basis of this experience. They know their own learning style; they look for familiar structures and cognates; they try to understand the main ideas of a text instead of going into details; they can deal with their own deficiencies, etc. (Hufeisen, 2001). However, learning an L3 that is cognate to L2 can give an additional advantage. For instance, learners of Polish (L3) with knowledge of Russian (L2) do not need to start from scratch, because they already know some common phenomena in the two languages, such as principles of conjugation. This is not the case for those who have studied German as L2, for example. They will notice the differences and common phenomena across the two languages, but these will appear less systematic. So, the contrastive language learning approach, applied to cognate languages, is likely to give the learner real advantages, as recent research has demonstrated. For example, English students with some knowledge of French were able to efficiently acquire reading skills in Romanian by focusing on lexical similarities and systematic differences across the two languages (Ciobanu, 2006).

EuroCom (Klein, URL) is a project associated with the idea of using one language, usually that which is most widely taught within the language family, as a basis for teaching its cognate languages. The main principle of the EuroCom method is the rejection of the simultaneous acquisition of all competences, in favour of concentrating on the teaching of receptive competences, especially reading. Unfortunately, no results have been reported for the EuroComSlav project, and the resources for the related EuroComRom (e.g., Pan-Roman vocabulary) are not particularly useful for Slavonic languages.

Nowadays language learning methodology focuses more than in the past on *authenticity* in contents, context, and task (Rüschoff, URL). Thus, the focus is on exploitating authentic materials even when dealing with tasks such as the acquisition of grammatical structures and lexical items. Accordng to this principle, learners should have the opportunity to discover language rules by themselves using digital materials related to their respective area of interest. Corpus work, particularly concordancing, is well-suited to Language for Specific Purposes (LSP) lessons, since it brings to light regularities in context, leading to the acquisition by the learner of large specialized schemata (Bernardini, 2004: 17). Moreover, it promotes an active and constructive learning process through authenticity in activities, such as finding information required for completing a task, analyzing it, finding solutions and adapting them to learners' needs.

However, none of the proposals based on DDL principles that we have examined goes beyond the use of concordances for deducing or interpreting word meanings from context, finding collocations and identifying typical patterns. DDL does not usually go beyond the word level and so does not reach text level phenomena, such as textual cohesion. Certainly, existing concordancing tools are not very instructive at the text level.

Our approach examines reading competence from the perspective of those reading processes activated while reading. Some of the processes operate at lower levels, that is: lexical access, syntactic parsing, semantic proposition formation and working memory activation. Higher-level processes are: text model of comprehension, situation model of reader interpretation, the use of background knowledge and inferencing, and executive control processes (Grabe & Stoller, 2002: 20 ff.). The lower-level processes imply automatized activities, whereas the higher-level processes involve meta-skills. Without underestimating the importance of the former set of skills, our approach accords greater importance to the latter. Thus we prefer to adapt the top-down model of reading rather than the bottom-up one in devisingthe best methods for teaching reading in L3. We consider the higher-level processes essentially as a complex of reading strategies performed at a text level and beyond.

3.Text classification using connectors

Given our focus on authentic text, we were faced with the problems of selecting texts for the classroom, classifying them and identifying their structure and genre. We therefore conducted a two-stage experiment on text collection and classification using *connectors*, that is, the "units" of conjunction.

We made at the outset the assumption that *conjunction*, as a type of text cohesion (Halliday & Matthiessen, 2004), is a marker of structure and genre that is highly relevant to foreign-language teaching (FLT) tasks. Functioning to mark semantic relations between parts of the text, conjunctive elements signal the logical text structure. Conjunction was selected as a primary textual cohesive device for two main reasons.

1. We assume that, of all cohesive devices, conjunction is significant for acquiring reading abilities effectively. The focus on conjunction responds to the call for text-focused applications in FLT.
2. Conjunctive elements are explicit enough to be captured with Natural Language Processing (NLP) methods.

Although Halliday & Hasan (1976) define the "units" of conjunction as *conjunctives*, *conjunctive adjuncts,* or *discourse adjuncts*, and later (Halliday & Matthiessen, 2004) even as *conjunctions*, we use the term connectors, because it is more consistent with the terminology known by language teachers and learners across languages. Also, the term conjunction is easily confused with a part-of-speech label.

Despite the awareness of the importance of using cohesive devises for text production in FLT, their usefulness for text reception has remained neglected, especially for East Slavonic languages.

During the first stage of the experiment, we compiled lists of connectors for each of our three languages: English, Russian and Ukrainian. Since for Russian and Ukrainian no classification of conjunctive relations is available, we derived them by translation from (Halliday & Matthiessen, 2004), which at the same time provided a basis for their classification by type of semantic function. We also collected them from academic grammars of Russian and Ukrainian, and later extracted them from our corpus. Our aim was to detect which connectors are the most significant and characteristic for marking certain semantico-logical relations, rather than to undertake a full, detailed categorisation of connectors. Currently we identify 343 connectors in 14 categories, listed in Table 1 and exemplified in Table 2.

TIME	ADDITION
ARGUMENT	REFERENCE
REASON	PURPOSE
RESULT	CONCESSION
CONDITION	COMPARISON
EXEMPLIFICATION / CLARIFICATION	SEQUENCE /CONCLUSION
ADVERSARIAL	OPINION

Table 1: Classification of connectors

	EXEMPLIFICATION / ИЛЛЮСТРИРОВАНИЕ
RUS	то есть, или, а именно, именно, как-то, то-бишь, как раз, точно, ровно, приблизительно, почти, в частности, к примеру сказать
UKR	тобто, а саме, себто, наприклад, а саме, на зразок, на кшалт
ENG	that is, in other words, I mean, to put it in another way, for instance/example, to illustrate
	CLARIFICATION / УТОЧНЕНИЕ
RUS	а точнее, буквально, в частности
UKR	скоріш(е), точніш(е), щоб бути точнішим/ою, буквально
ENG	or rather, at least, to be precise, more especially

Table 2: Example connectors.

These relations formed the basis for defining a set of text organization patterns.

At the same time, comparable texts on the topic of the Warsaw Pact were collected for all three languages – 303 texts in English, 156 in Russian, 114 in Ukrainian, -- using the methodology proposed in (Baroni & Bernardini, 2004). These were then filtered by their length (maximum

1,500 words) and according to the presence of connectors. We discarded all texts without a single connector (leaving 54 English, 34 Russian and 84 Ukrainian texts), then calculated the frequency and relative frequency of each connector / each class. The end result was a *conjunction profile* for each text.

By looking at the configuration of connector types in each text, it was possible to select texts according to the closure / openness of their ‘parent’ register. The distinction between open and closed register is based on (Halliday & Hasan, 1997: 39): “The category of register will vary, from something that is closed and limited to something that is relatively free and open-ended. That is to say, there are certain registers in which the total number of possible meanings is fixed and finite and may be quite small; whereas in others, the range of the discourse is much less constrained.” For instance, the number of meanings in an instruction will be considerably smaller, than in everyday conversation. In our corpus the restricted register was represented by, among others, encyclopaedia articles, and more open-ended registers by interviews and commentaries.

The text filter based on the presence and distribution of connectors proved to be able to remove non-cohesive texts among all collected with an accuracy of 100%. Moreover, it proved possible to select texts according to the register using the conjunction profile:

1. Texts having one or two classes of connectors with low relative frequency (<0.1%) proved not useful, e. g. lists.
2. Texts having one or two classes with higher relative frequency (>0.1%) as a rule belong to a restricted register.
3. Texts having more than two classes, some with high relative frequency, fall clearly within more open-ended registers.

The first stage of the experiment suggested that texts can be reliably classified into ‘open’ and ‘closed’ register using only connectors as features. In order to verify this hypothesis, we carried out a further experiment using a machine learning toolkit. We selected 48 Ukrainian texts and manually classified them into ‘information’ and ‘opinion’ categories, which correspond to our distinction between open and closed registers.

These two text types have very different uses in the teaching of reading, and making such a distinction automatically would be beneficial for many different CALL applications that use automatically downloaded Internet corpora.

For the training corpus, which comprised the 48 texts, we counted for each text the number of connectors in each of the classes in Table 1, and used these counts as features for predicting whether the text belongs to the ‘information’ or ‘opinion’ category. We used the Weka implementation of SVM (Witten & Frank, 2005) with 10-fold cross-validation to estimate the accuracy of the classifier. Table 3 presents the results of this evaluation.

The classification is reliable (achieving an F-measure of 0.86 and 0.93 for the two classes), even despite the fact that we used a relatively small dataset for training the classifier. Predictably, the most important class of connectors for differentiation between two text classes is the OPINION class of connectors, (e. g. *in my opinion, as well known*) with a weight of 1.4088, followed by the classes of PURPOSE (1.391) and COMPARSION (1.0125).

Instance classification		
Correctly classified instances	44	91.67 %
Incorrectly classified instances	4	8.33 %
Total number of instances	48	

Confusion matrix		
<-- classified as	a	b
a = information	31	1
b = opinion	3	13

Detailed accuracy by class

Class	TP rate	FP rate	P	R	F
inf’mation	0.969	0.188	0.912	0.969	0.939
opinion	0.813	0.031	0.929	0.813	0.867

Table 3: Weka accuracy in classifying texts

4.Conclusions and future work

Our experiments show that CALL applications can benefit from our corpus-based tools that use linguistic models of textual cohesion. Such models reliably predict text types that are useful for data-driven learning, e.g., for automatically selecting texts, and can greatly enhance the efficiency of teaching reading in foreign languages.

Future work will include developing a methodology and a set of computer-assisted tools for L3 teaching. The tool will align comparable texts in Ukrainian, Russian and English by their topic and text type. They will belong to one specific domain.

We are in the process of designing a course which will take into account the results of current research in reading in a foreign language and in text linguistics. The course will give an overview of the differences between Russian and Ukrainian grammar and lexicon. This will require development of core NLP resources for Ukrainian, which are still not available in the public domain. These NLP tools will include a Ukrainian part-of-speech tagger and a lemmatizer, tools for discovering cognates in the two Slavonic languages automatically. These tools will be used as a basis for systemic acquisition of Ukrainian grammar. This will be achieved via discovery learning, by contrasting linguistic categories and features found in cognates with the specific characteristics of Ukrainian. The text classification will be based on the set and principles outlined in (Sharoff, 2008), while developing new methods for collecting up-to-date corpora and texts of controlled length.

To test our hypothesis about the dominant role of the top-down model for reading comprehension in our constellation of learning situation and task, an interactive on-line system will be created for supporting the proposed methodology. The methodology will be tested with students during a course. designed to accommodate autonomous learning.

References

Baroni, M., & S. Bernardini. (2004). *BootCaT: Bootstrapping corpora and terms from the web.* LREC, Lisbon.

Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (Vol. 12, pp. 15-36). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Ciobanu, D. (2006). *Acquiring Reading Skills in a Foreign Language in a Multilingual, Corpus-Based Environment.* PhD Thesis. University of Leeds, Leeds.

Ciobanu, D., T. Hartley & S. Sharoff. (2006) *Using Richly Annotated Trilingual Language Resources for Acquiring Reading Skills in a Foreign Language*. Procs. LREC, Genoa.

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

Halliday, M. A. K., & C. Matthiessen. (2004). *An introduction to functional grammar* (3 ed.). London: Arnold.

Halliday, M. A. K., & R. Hasan. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & R. Hasan. (1997). *Language, context, and text: aspects of language in a social-semiotic perspective*. Geelong: Deakin University.

Hufeisen, B. (2001). Deutsch als Tertiärsprache - German as a third language. In G. Helbig & L. Götze & G. Henrici & H.-J. Krumm (Eds.), *Deutsch als Fremdsprache. Ein internationales Handbuch* (Vol. 1, pp. 648-653). Berlin, New York: Walter de Gruyter.

Klein, H. G. *EuroCom Website.* Accessed 31-03-2008 http://www.eurocomresearch.net/

Rüschoff, B. *Data-Driven Learning (DDL): the idea by Bernd Rüschoff.* Accessed 31-03-2008: http://www.ecml.at/projects/voll/rationale_and_help/booklets/resources/index_ddl.htm

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.

Sharoff, S. (2008). In the garden and in the jungle: comparing genres in the BNC and Internet. To appear in A. Mehler, S. Sharoff, M. Santini, G. Rehm (Eds.), *Genres on the Web*. Springer.

Witten, I. & E. Frank (2005). Data Mining: Practical machine learning tools and techniques. San Francisco: Morgan Kaufmann. Available [2008, 31-03-2008]: http://www.cs.waikato.ac.nz/~ml/weka/book.html