PORTA LINGUARUM 6, junio 2006 pp. 33-39

Computer-assisted early inclusion of authentic Slavic materials

DANKO SIPKA
Department of Languages and Literatures
Arizona State University

Received: 10-2-06 / version accepted: 12-3-06
ISSN: 1697-7467

ABSTRACT : The author discusses Bosnian/Croatian/Serbian, Polish, and Russian text taggers, available at **http://www.asusilc.net/cgi-bin/newtepajgu.pl**. The taggers allow the user to paste in a text, copied from an on-line source and have it tagged with English glosses. They furthermore offer the option of displaying the full inflection of each inflected word form in the text. The taggers are designed with an eye toward securing optimum authenticity and learner autonomy. The present paper summarizes theoretical background of this project and its achievements hitherto. It furthermore identifies major problem areas of this project and outlines its envisaged development hence.
Key words: taggers, CALL, authentic materials, slave languages.

RESUMEN: El autor trata las expresiones en lengua bosnia, croata, servia, polaca y rusa, disponibles en **http://www.asusilc.net/cgi-bin/newtepajgu.pl.** Estas frases y párrafos ya elaborados permiten que los usuarios los'"peguen" en un texto, los copien de una fuente "on-line" y visualizarlos traducidos momentáneamente en lengua inglesa. También permiten la opción de mostrar la inflexión de cada palabra y sus morfemas. Estos textos gozan de una autenticidad óptima y favorecen el aprendizaje autónomo. Este artículo se propone resumir el marco teórico de este proyecto didáctico y dar a conocer algunos de sus logros hasta el momento presente. Además, identifica alguna áreas problemáticas del proyecto y trata de dar soluciones para su desarrollo posterior.
Palabras clave: circunloquios, CALL, materiales auténticos, lenguas eslavas.

1. INTRODUCTION

One of the hallmarks of the early twenty-first century is the shifting locus of educational training from the traditional fixed classroom to the Internet and real-life immersion. Language instruction is not an exception in this respect. There is a demand for on-line and immersion language learning which will enable the student to act as an independent performer of various tasks in the target language (e.g., understanding authentic materials, engaging in on-line and face-to-face interactions, capturing cultural differences, etc.). In contrast to these needs, most available less commonly taught languages textbooks remain limited to traditional in-class instruction where rote learning detached from real life and learners as objects rather than subjects is perpetuated. Similarly, courses of foreign languages as a rule do not incorporate

immersion or task-based e-learning. In effect, learning outcomes are reduced to passive knowledge of grammatical structures and vocabulary, rather than to linguistic and cultural literacy in multiple genres of the target language. This unfortunate state of the affairs is a frequent subject of concern in the academe (see for example Brecht and Rivers, 2002). See also: bibliography.

Despite the widespread I-just-want-to-speak student attitude, the most valuable professional skill to be acquired in Slavic language classrooms is the ability to understand authentic written and spoken texts in the target language. The importance of reading and listening skills in a real-life setting is obvious in a range of professions – from scholars of history, sociology, economics, etc. to intelligence officers. It is therefore imperative that authentic materials are included in the curriculum as early as possible. However, attempting to address this imperative leads to the following dilemma. On the one hand, students with limited command of the vocabulary cannot be expected to process "raw" authentic texts as constant dictionary lookup would be overly time consuming and frustrating. On the other hand, instructors do not command sufficient financial and temporal resources to manually gloss these texts with English equivalents. Being able to confront authentic materials early in the instructional process bears a two-fold importance. First, the student acquires the sense of achievement, the feeling that she/he can utilize the language in a sensible manner, which in and of itself wields a beneficial effect on the cognitive and affective factors in language acquisition. Secondly, the early inclusion of authentic materials creates a solid cognitive basis for later phases of the process in which authentic texts become ineluctable. Both these needs are broadly recognized in the current movement for foreign language learner autonomy.

The body of literature addressing authentic materials and learner autonomy in foreign language classroom is voluminous and diversified. While key concepts and their definitions are far from being commonly accepted, it is important to establish a conceptual map by defining several key concepts as they pertain to the present research. In an attempt to emphasize the complexity of criteria for authenticity the following definition of authenticity was adopted.

"Authenticity is a factor of the:

1. Provenance and authorship of the text.
2. Original communicative and socio-cultural purpose of the text.
3. Original context (e. g. its source, socio-cultural context) of the text.
4. Learning activity engendered by the text.
5. Learners' perceptions of and attitudes to, the text and the activity pertaining to it." (Mishan, 2004:18)

It is of paramount importance to note that not the text per se but rather its socio-cultural and pedagogical interactions define the notion of authenticity. Another concept that deserves our attention here is that of learner autonomy, discussed recently in Mishan (2004:7): "Over the past thirty-odd years there has been a gradual shift of preoccupation in the field from teaching to learning and thence to the learner. [...] These changes reflect the recognition that it is the learner who stands at the centre of - and ultimately controls - the learning process."

Finally, one should bear in mind the evidence suggesting the need for instructional intervention in foreign language exposure. In a recent study, Jiang (2004: 121) concludes:

"The results of the present study and the previous ones reviewed earlier show that natural exposure alone may not provide enough impetus for semantic restructuring and development. [...] I'd like to argue in this context that instructional intervention has an important role to play in helping learners overcome plateaus in semantic development".

In a related study, Ridder (2003) concludes that there seems to be a link between electronic glosses on the one hand and text comprehension as well as incidental vocabulary learning, on the other hand. However, effectiveness of glossing is highly dependent on reader's learning style.

Slavic text tagger projects are an attempt to resolve the dilemma between pressing instructional needs and limited resources. Providing tools which facilitate the comprehension of written texts (newspaper articles, short stories, public and corporate web sites, etc.), meets the aforementioned instructional needs with little or no resources required by the instructors. More information about the general framework of the project can be found in Sipka (2004). The present project is meant to complement the existing immersion materials, such as those listed at the UCLA Language Materials Project (**http://www.lmp.ucla.edu/SampleLessons.aspx**). Concurrently, this project continues the established tradition of using corpora in language learning. An informative review of this tradition can be found in chapter four (The Use of Corpora in Language Studies) of McEnery and Wilson (2001). One should also consult data-driven studies within this tradition carried out at the Centre for English Corpus Linguistics of the Université catholique de Louvain (see: **http://cecl.fltr.ucl.ac.be/**) and reported in the publications such as Granger *at al.* (2002) and Granger and Tribble (1998).

2. DESIGN OF THE TAGGERS

Taggers for Bosnian/Croatian/Serbian (BCS), Polish and Russian are generally available and located at **http://www.asusilc.net/cgi-bin/newtepajgu.pl.** The taggers are a part of a wider project titled Learner-centered Task-oriented Language Instruction, presented at **http://www.asusilc.net/lctli**. Each of the three taggers accept electronic texts in various formats (UTF-8, cp-1250, cp-1251, etc.), either typed into the text window or copied from an internet page and pasted into the window.

The taggers return the text tagged with the English glosses as can be seen at **http://www.asusilc.net/lctli/exbcs.htm** (BCS), **http://www.asusilc.net/lctli/expol.htm** (Polish), **http://www.asusilc.net/lctli/exrus.htm** (Russian).

By clicking at any of the underlined word forms in the text, the user can get their respective English gloss. Thus, clicking at the BCS word *sahrana* will yield the following gloss: *sahrana,e f [I] funeral n, sepulture n, interment n, inhumation*. Pressing the I (inflection) button will expand the word *sahrana* in all its forms (Nominative Singular *sahrana*, Genitive Singular *sahrane*, etc.). The design of the taggers is intended to facilitate the cumbersome and time-consuming process of looking up the English equivalents and determining the morphological category of the word forms in the texts.

The following technologies were utilized to implement the taggers. The entire knowledge base, with dictionary forms, inflectional forms and the English equivalents is stored into a relational database in MySQL. Perl scripts with HTML forms as their GUI are used to query the database and tag the text, while the resulting tagged text is implemented in a form of HTML, DHTML with a limited use of Java Script. Central to this design was the idea that

all operations are performed serverside, which stipulates minimum requirements on the part of the user, coupled with the transferability of the resulting HTML page with the tagged text.

The taggers accept the input text in various code pages (Unicode, Windows, ISO, etc.) and the output text is always rendered in Unicode. It is important to note that the BCS tagger can handle both Latin and Cyrillic script.

The resulting tagged page can be saved and post-edited. An example of a post-edited text created using a previous version of the BCS tagger and incorporated into the BCS 201-202 courses can be found at: **http://www.public.asu.edu/~dsipka/bcs202a1.html**. The background about methodology and technology in on-line delivery of BCS can be found in Sipka (2003).

Presently, the most advanced is the BCS tagger (the access is available at: **http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=sr**), which covers approximately 95% of an average newspaper text. The Polish tagger (**http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=pl**) covers some 85% of an average newspaper text, while the Russian tagger (**http://www.asusilc.net/cgi-bin/newtepajgu.pl?lang=ru**) covers over 60%.

In the present version of the taggers, the text is tagged automatically, and all possible tags are listed. Thus if we have the BCS form *je*, which can be either accusative or genitive singular of the personal pronoun *ona* 'she' and the third person singular of the verb *biti* 'to be', both these glosses are listed. This manner of tagging replaced the previous version of the tagger, in which the text could be tagged automatically or interactively. In the former case the most frequent solution was deployed in case of ambiguous forms (e.g., in the example above, the verbal meaning would be selected). The interactive manner of tagging was prompting the user to resolve all cases of ambiguous forms (e.g., to tell if the form *je* found in the text belongs to the verb *biti* 'to be' or to the pronoun *ona* 'she'). However, it turned out in the course of longitudinal testing that manual tagging is overly time-consuming and annoying while automatic selection of only one tag bears a risk of being inaccurate. The present solution allows swift tagging and concurrently provides the user with the opportunity of selecting the right equivalent while using the tagged text.

An average newspaper text of 350 word forms is tagged in six seconds. Other statistical data for this server (e.g., number of hits per day for any given month) can be accessed at **http://www.asusilc.net/stats.**

3. PREPARATORY, ACCOMPANYING, AND FOLLOW-UP ACTIVITIES

The activities centered on the tagged text are designed to secure maximum authenticity, learner autonomy and development, as described in the first section of this paper.

Going back to the criteria for authenticity, we can see that the first three (provenance and authorship; purpose; context) pertain to the text itself, while the remaining two (learner activity; learners' perceptions and attitudes) relate to its user.

Assurance of the first three criteria is the preserve of preparatory activities. In a situation where the student can paste any material found on the Web, it is of paramount importance to guide the students toward the texts which meet all three criteria. Preparatory activities encompass in-class explorations of available media sources, their sections, various genres,

etc. In addition, students are directed to the texts appropriate to their proficiency level and targeted learning outcomes.

Learner-related authenticity is secured in all three phases of student engagement (pre-reading, reading, and post-reading). In the pre-reading phase the students are steered toward the materials concurrently attuned to intended learning outcomes and their own personal interests. The reading phase activities always include an attempt to capture the import of the text, the way fully proficient readers would read their Sunday morning newspaper. This authentic task is then a "staging area" for other, more pedagogically related activities, such as focusing on the form (inflections, collocations, etc.). The post-reading phase typically includes summarizing the content of the reading and discussing it with one's peers, the way fully proficient reader would retell a newspaper story to his/her interlocutor and discuss it with him/her.

While some non-authentic tasks (such as finding the accusative forms and explaining why the author used them) may accompany the aforementioned activities, there is always an authentic core aimed at securing all five criteria for authenticity.

It is a matter of course that the learner acts autonomously while using the tagger. Being aided by the tagger and not forced into a frequent dictionary lookup, he/she is far more likely to stay with the task, which in turn is more likely to produce the desired learning outcomes.

4. MAJOR CHALLENGES

The major challenge currently addressed within the Slavic text taggers project is that of appending and amending the knowledge bases. At present, the team members are working on formatted text files, which are then transferred into the databases using PERL scripts.

Given that this solution does not allow computational lexicographers and grammarians to immediately introduce a change and see its result, the development of a knowledge base authoring tool is underway. The completion of this tool is envisaged for the last quarter of 2005 and in its completed form it will allow addition of new items as well as modification of the inflectional forms and the English equivalents. Central to the development of this resource is to allow concurrent use of various updating techniques – automatic (using the available resources) and manual, modification of one peculiar form of inflection and assigning a completely different paradigm, etc.

Two major techniques are being deployed in appending and amending the knowledge bases. First, the databases are filtered using available monolingual and bilingual lexical lists or dictionaries and the missing entries are added. Second, the tagger itself is used as an important tool of testing the comprehensiveness and accurateness of the database. As can be seen in the Appendix, if a word form is not lemmatized, it will not be underlined, if it is lemmatized yet its English equivalent is missing, it will not be in the bold type. Those forms which are lemmatized and equipped with the English equivalent will be bold and underlined. Testing the taggers with the texts from the most popular newspapers is used to train the knowledge bases.

Inasmuch as the three languages exhibit a varied degree of development, the concrete tasks performed on the databases expose considerable differences. In the case of BCS, some low-frequency lexemes and irregular forms are still being added to the knowledge base. At

present, the process is geared toward appending the knowledge base with the items from a 230,000-entry word list, which served as the bases for Sipka (2002). Even upon full inclusion of the aforementioned list, there will still be some 1-2% unrecognized forms, owing primarily to the creativity of the journalists and other authors. For example, the negative particle *ne-* can be added to practically any noun or adjective, as in *neodgovaraju_i* 'unfitting, unsuitable' in the Appendix below. This problem will be tackled by tagging parts of the word form if the tagging of the form in its entirety fails. Obviously, proper names will not be tagged at all.

The problems being solved in the Polish and Russian tagger are quite different to those in the BCS tagger. Here, the goal is to form a solid foundation of approximately 80,000 lexemes (i.e., canonic forms) with their corresponding equivalents and inflections. The synergy of tagger training of the knowledge base and its appending using the available electronic resources is deployed here in a manner akin to the BCS tagger project.

5. FURTHER DEVELOPMENT

The goal of the project is to reach 95% converge for the Polish and Russian taggers and 98% coverage for the BCS tagger by October 2006. All activities in the intervening period will be subordinated to that overarching goal. A further step will include the expansion of the resources to other languages beyond the Slavic linguistic realm. Finally, a number of spin-off projects are envisaged. Most notably, a tool will be created for semi-automatic collections of neologisms. A net-bot will be implemented to randomly search and tag a predetermined set of web pages end extract unattested lexemes along with their contexts for a subsequent human lexicographic treatment.

6. REFERENCES

Bogaards, Paul (ed.) (2004). *Vocabulary in a Second Language. Selection, Acquisition, and Testing.* Philadelphia, PA, USA: John Benjamins Publishing Company.

Brecht, R. D. and W. P. Rivers (2002). "The Language Crisis in the United States: Language, National Security and the Federal Role", in: Baker, Stephen J (ed.) *Language Policy: Lessons from Global Models*, Monterey: MIIS, 76-90.

Byram, M (ed.) (2001). *Developing Intercultural Competence in Practice.* Clevedon, GBR: Multilingual Matters Limited.

Ellis, R (ed.) (2005). *Planning and Task Performance in a Second Language.* Philadelphia, PA, USA: John Benjamins Publishing Company.

Forsyth, I. (2001). *Teaching and learning materials and the Internet*. London: Kogan Page.

Granger, S. and C. Tribble (1998). "Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning". in: Granger (ed.) *Learner English on Computer*, London: Addison Welsey Longman, 199-209.

Granger, S.,Hung,J. and Petch-Tyson, S. (eds.) (2002). *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam and Philadelphia: John Benjamins.

Jiang, N. (2004). "Semantic transfer and development in adult L2 vocabulary acquisition". In Paul Bogaards & Batia Laufer (eds.) *Vocabulary in a Second Language: Selection, acquisition, and testing*, 101-126.

McEnery, A.M. and Wilson, A. (2001). *Corpus Linguistics (second edition)*, Edinburgh: Edinburgh University Press.
Mishan, F. (2004). *Designing Authenticity into Language Learning Materials.* Bristol, GBR: Intellect Books.
Ridder, Isabelle De (2003). *Reading from the Screen in a Second Language*, Antwerpen: Garant.
Sanderson, P. (1999). *Using newspapers in the classroom*. Cambridge: Cambridge University Press.
Sipka, D. (2003). "On-line Delivery for Serbo-Croatian (Bosniac, Croatian, Serbian)", *Journal of the NCOLCTL*, Volume 1, 95-118.
Sipka, D. (2004). "Content-centered resources for the West Balkans" In R. Jourdenais, & S. Springer (eds.). *Content, tasks and projects in the language classroom: 2004 conference proceedings.* Monterey, CA: Monterey Institute of International Studies, 123-129.
Sipka, D. (2002). *Enigmatski Glosar. Dio 1. Osnovni oblici*, Beograd: Alma (2002), 1-1315.

7. APPENDIX

Appendix 1: Screen caption of the BCS tagger