Using corpora to find relevant examples of Japanese honorific forms in order to generate exercises for intermediate learners of Japanese.

Keywords: corpus linguistics, Japanese linguistics, Japanese honorifics

Abstract: *This article focuses on the importance of the relevant use of corpora even in domains like automatic sentence generation. We will first present our study, and then show that even if the use of huge existing corpora is indispensable, it is not entirely satisfactory to meet all the requirements of our domains. To come to an end we will propose an approach to avoid the shortcomings brought about by an only statistical corpus-oriented analysis within the framework of our specific field of application which is sentence generation to help intermediate learners of Japanese to reach higher level.*

Our study aims at helping intermediate students to improve their knowledge outside an usual in praesentia learning environment by proposing them additional exercises automatically generated.
The task of learners will be to rebuild the sentences originally generated. Some researchers have showed that it is possible help beginners to learn basic syntactic structures, by offering them an electronic version of phrasebooks augmented with exercise generators (Zock, Lapalme 2010). For intermediate learners, semantic and syntactic information are much more intertwined. Furthermore pragmatic and semantic considerations could lead intermediate learners to build up complex sentences or to use a much larger set of rules. The linguistic politeness of the Japanese language, which is more often referred as honorific language, is one of the clearest examples of this complex linguistic phenomenon. The linguistic honorific system is narrowly linked to the Japanese socio-cultural landscape. It can be measured either vertically (superior-inferior relationship, socioeconomic class, age, etc…) or horizontally (giving-receiving relationship, in-group out-group relationship) (Wlodarczyck, 1996) (Wetzel, 2004). To choose the most appropriate linguistic constructions, the participants of the speech act must be able not only to position themselves in connection with their conversational partners but also with the actors of the utterance who are not necessarily present during the process of the speech act. These constraints make the use of the honorific system extremely difficult for non native speakers and limit their opportunity of using Japanese in professional communication.
However, nowadays many Japanese web sites offer sentence patterns of honorific language to guide native speakers in using honorific forms properly. That is why we assume that it could be possible to partly formalize some of the usages of the honorific forms to generate sentences in connection with the context of use.
Since the early 80ies researches have been conducted for this purpose within the framework of grammar formalisms. For instance, we can mention S. Tanaka (Tanaka 1983) who attempted to give a

computational description of the honorifics based on Shizuo Mizutani's work about a systemic description of the Japanese, R. Sugimura (Sugimura 1986) and his model of honorific expressions in semantic situation as well as A. Wlodarczyck (Wlodarczyk, 1988) in his attempt of PROLOG implementation of the honorific systems, and more recently M. Siegel (Siegel 2000) and her HPSG approach of Japanese honorification. We have used some of these elements as well as some grammatical books (Schikowski, 2008) (Shimamori, 2001), (Tsujimura, 2005), (Terrya, 2007) to build up our first rules of sentence generation.

As the use of honorifics is not only linked to the situation of communication but also constantly evolving, we will restrict our work to the generation of sentence models that intermediate learners can use in the most frequent situations and help them to acquire linguistic reflexes. Thus it will be afterwards easier for the learners to do local adjustments when speaking to real conversational partners. It is however essential that our system should not generate improper sentences, that is why we regard the use of corpora as indispensable to verify the representativeness of the generated sentences. We have also chosen to have these sentences validated by native speaking Japanese teachers of Japanese. The roles of these teachers are very significant. Indeed, in addition to the validation of the syntactic and semantic correctness of the sentence, they could also assess if it is worth learning or not.

To meet our requirement we have chosen to consult two on-line corpora: the Balanced Corpus of Contemporary Written Japanese (BCCWJ) and the corpus associated to the software Sagace. (Blin 2009) The BCCWJ is a compilation of about 100 million words produced by the National Institute for Japanese Language and Linguistics (NINJAL). This corpus includes a various range of text genres from books, magazines, newspapers to governmental texts, bulletin boards and blogs and so on. Sagace is a tool designed to retrieve and extract linguistic patterns from a corpus which consists of more than 20 million phrases from books, newspapers, texts of law, dictionaries, patents, discussion forums, etc. These corpora must validate the syntactical constructions generated by our system (downstream-validation) and help us to choose the forms that are the most frequently used. However, the public version of the Balanced Corpus of Contemporary Written Japanese does not enable the user to apply extended regular expressions, whereas Sagace only enables researches based on strings of characters, even regular expressions identifying the syntactic categories can be used. These two tools were used in order to give a first validation or to have an overview of the representativeness of the pattern we are looking for in a huge set of texts. But the patterns which are retrieved by the software give only information about the utterance which could sometimes describe the whole sentence but rarely gives direct information about the participants of the speech act. Moreover the tools rarely provide detailed

analysis, because text-annotations if any, rarely meet specific needs. That is why we have decided to build up a little sample of corpus in parallel, which is composed of the retrieval of recent texts from websites.

The specific problems linked to the Japanese language for corpus analysis are well known: multiple encodings, the absence of separation markers between words and the poor number of Sino-Japanese ideograms, which made it difficult to recognize the parts of speech as well as the forms in a given text. Since automatic word segmentation performed by statistical parsers like Chasen or MeCab could be mistaken, we have chosen to insert our small training corpus into the Nooj tool. Nooj is a parser using graphs (Silberztein, 2007). It enables direct annotation of corpora for specific purpose without word segmentation. In addition to linguistic resources such as a dictionary, we provided this tool with a small set of representative texts which had been selected from the Internet. The selection was based on the compliance with the results previously obtained through Sagace and the BCCWJ according to the representativeness of the pattern, so that Internet could enable us to examine text in its full context.

The web sites in which honorifics are used are mainly requests for advice from internet-users about the use of honorifics, advice about the use of honorifics in business communication, journalistic transcriptions of the celebrities' interviews, official meeting minutes, and commercial sites and blogs. We rarely obtained texts which deal with transcriptions of job-interviews, or teacher-student relationship and we think it would be extremely difficult to have them due to confidentiality. However, we hypothesize that the rules of the honorifics could be enlarged to equivalent social groups. This hypothesis enables us to simplify the different situations of communication that are encountered in the real world, which is coherent with our context of Japanese learning. However, in every case an automatic treatment is not sufficient to determine the context surrounding the utterance, therefore a manual semantic analysis (Rastier, 1995) is unavoidable. This constraint dramatically limits the number of texts that could be analyzed and as a result the size of potential corpora. Nevertheless we consider that a manual pretreatment enables more precise formal descriptions.

We assume that an analysis only based on statistical data from huge general corpora is not sufficient to determine situation of communications that is linked to the utterance. However it enables our results to be partially validated. Afterwards we were led to use an open research in the internet to obtain more contextualized information. So we may risk to generalize particular cases of use if the results obtained is close to what we have already obtained from our formal rules.

References

Biber D. et al. 1998. *Corpus linguistics Investigating language structure and use,* Cambridge University Press

Blin R. 2009. *SAGACE; Analyseur de corpus pour langues non flexionnelles,* TALN 2009, ATALA.

Rastier, F. (dir.) 1995 *L'analyse thématique des données textuelles,* Paris : Didier, 1995 p. 223-249.

Schikowski, R. 2008 *Skript zum Strukturkurs Japanisch*. LMU Munich, Institute of General and Typological Linguistics, winter term 2007/08

Shimamori, R. 2001 *Grammaire japonaise systématique,* volume II - Edition Jean Maisonneuve Paris 12ème édition, 2001

Siegel M. 2000. *Japanese Honorification in an HPSG Framework,* Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation. p. 289-300.

Silberztein M. et al. (Ed.). 2007. *Formaliser les langues avec l'ordinateur : de INTEX à NooJ***.** Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté, Besançon.

Sugimura, R. **1986.** *Japanese honorifics and Situation Semantics,* International Conference on Computational Linguistics COLONG. p. 507-510.

Tanaka S. et al. 1983. *Keigo wo totonoeru (Treatment of the honorific form), in «Asakura Nihongo Shin-Kôza», vol. 5,Asakura Shoten, Tokyo..*

Terrya, K. 2007. *Interpersonal grammar of Japanese* in A Systemic functional grammar of Japanese**,** Vol 2, Ch 4**,** p135-205.

Tsujimura, N. (ed.) 2005. *Japanese Linguistics Vol II syntax and Semantics Vol III Pragmatics, Sociolinguistics and language contact* Ed Routlege London

Wetzel, P. 2004. *Keigo in modern Japan from Meiji to the present,* University of Hawai press

Wlodarczyk, A. 1988. *Les traits pertinents du système honorifique japonais - une tentative d'implémentation en Prolog,* communication au 5e Congrès de l'EAJS, Durham - 1988, in «European Studies in Japanese Linguistics 1988-90», Lone Publications, London, 1991 (pp. 127-150).

Wlodarczyk, A. 1996. *Politesse et Personne – Le japonais face aux langues occidentales*. Editions L'harmattan.

Wlodarczyk, A. 2007. *Towards a Unified Treatment of Linguistic - Person and Respect - Identification* Japanese Linguistics – European Chapter Tokyo.

Zock M. / Lapalme G. 2010 *A Generic Tool for Creating and Using Multilingual Phrasebooks.* Natural Language Processing and Cognitive Science.