The aim of register analysis is to investigate the link between the linguistic expression and the social situation, with a view towards explaining that link [3: 3–12]. In the same manner, the analysis of the register of electronic chatting aims at exploring and explaining the following:

- how the characteristics of the situation in which CMC takes place change the language and to what extent;
- what linguistic and extralinguistic features make this register different from other registers;
- is there a connection between extralinguistic features (e.g. lack of face to face contact, keyboard and computer screen as the only ways of producing and receiving the information) and linguistic features (e.g. typos, innovativeness and brevity of expression).

Finally, the aim of register analysis is to identify and interpret the generalizations concerning register variations, as well as to systematically list linguistic and extralinguistic features of the register [2: 29–30]. A point of importance here is that the linguistic features must be interpreted not only in relation to the extralinguistic features of text production, but also in relation to other linguistic features, because the isolated analysis of a linguistic feature may lead to wrong conclusions. Thus, the use of personal pronouns of the first and the second persons singular, direct questions and imperatives indicate interactivity, without disclosing if they are found in a written or a spoken text, while abbreviated forms and self-corrections point to a spontaneous spoken discourse.

4. Conclusion

After presenting a number of reasons for compiling an electronic annotated CMC corpus, it must be said that this is a task that should be done in as many languages as possible. This is because the interlinguistic comparison and contrasting reveals explicit similarities and differences between two or more languages and may draw attention to phenomena that would otherwise be missed. Also, that kind of analysis would reveal any linguistic and cultural differences in one register across two or more languages, thus contributing to a better understanding of other languages and cultures.

References

1. Biber, D. (1988). Variation across Speech and Writing. Cambridge University Press. Cambridge.
2. Biber, D. (1995). Dimensions of Register Variation. A Cross-Linguistic Comparison. Cambridge University Press. Cambridge.
3. Biber, D. & Finegan, E. (1994). "Situating Register in Sociolinguistics". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (3-12).
4. Crystal, D. (2001). Language and the Internet. CUP. Cambridge.
5. Danet, B.(2001). Cyberpl@y. Berg. Oxford, New York.
6. Ferguson, Ch. E. (1994). "Dialect, Register and Genre: Working Assumptions about Conventionalization". In: Biber, D. & Finegan, E. ed. Sociolingustic Perspectives on Register. Oxford University Press. Oxford. (15–30).
7. Hajmz, D. (1980). Etnografija komunikacije. [Foundations in Sociolinguistics. An Ethnographic Approach.] Beogradski izdavačko-grafički zavod, XX vek. Beograd.
8. Hine, C. (2000). Virtual Ethnography. SAGE Publications. London, Thousand Oaks, New Delhi.
9. Ooi, Vincent B. Y. (2000). "Aspects of Computer-Mediated Communication for Research in Corpus Linguistics". In: Peters, P., Collins, P. & Smith, A. ed. Language and Computers, New Frontiers of Corpus Research. Papers from the Twenty First International Conference on English Language Research on Computerized Corpora Sydney 2000. (91–104).

Р. Ющенко, В. Гудзенко*
Институт кибернетики им. В. М. Глушкова НАН Украины (Киев)
УДК 81’322.33

ПРИМЕНЕНИЕ ПАРАЛЛЕЛЬНЫХ КОМПЬЮТЕРОВ ДЛЯ АВТОМАТИЗАЦИИ МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ

Морфологическое аннотирование Национального корпуса предполагает значительные усилия, учитывая объем текстов, который составляет сотни миллионов слов. Большую часть рутины, однако, можно избежать, подключив к работе современные компьютерные технологии. Задача морфологической разметки поддается частичной формализации, а следовательно, процесс тегирования можно автоматизировать. В рамках этих работ Институтом украинского языка были

разработаны электронные морфологические словари словоформ украинского языка. Используя их, для большинства слов можно поставить в соответствие его морфологический разбор.

Каждый словарь образуется комбинацией части речи с ее словоформами [1]. Далее в таблицах приведены возможные словоформы частей речи. Здесь использованы условные обозначения для словоформ: род (m – мужской, f – женский, n – средний), число (p – единственное, s – множественное, t – pluralia tantum для существительных только множественного числа, d – двойственное), склонение (n – именительный, g – родительный, d – дательный, a – винительный, i – творительный, l – предложный, v – призывной), степень сравнения (1 – высшая, 2 – наивысшая), вид (f – совершенный, c – несовершенный, d – двувидовая форма), наклонение (a – изъявительное, m – повелительное, j – сослагательное), время (p – настоящее, t – прошедшее, u – будущее), лицо (1 – первое, 2 – второе, 3 – третье), состояние (a – активное, v – пассивное). Дефис в ячейке таблицы указывает на неопределенную форму соответствующей части речи.

Словоформа	Существительное	Прилагательное	Числительное
Род	m, f, n, -	m, f, n, -	m, f, n, -
Число	s, p, t, d	s, p, -	s, p, -
Склонение	n, g, d, a, i, l, v	n, g, d, a, i, l, v	n, g, d, a, i, l, -
Степень сравнения		1, 2	

Словоформа	Порядковое числительное и местоимение	Глагол	Инфинитив
Род	m, f, n, -	m, f, n, -	
Число	n, g, d, a, i, l, -		
Вид		r, c	r, c, d
Наклонение		a, j, m, z	
Время		p, t, u, c	
Лицо		1, 2, 3, -	

Словоформа	Причастие	Деепричастие	Наречие
Род	m, f, n, -	m, f, n, -	
Число	s, p, -		
Склонение	n, g, d, a, i, l, -		
Степень сравнения			1, 2
Вид	r, c	r, c	
Время	p, t		
Состояние	A, v		

Кроме вышеуказанных, в состав комплекса включены отдельные словари для союзов, предлогов и междометий.

Использование словарей для автоматический морфологической разметки приводит к проблеме эффективности разметки. Объем словарей составляет порядка миллиона словоформ. Учитывая, что для построения корпуса необходимо обработать сотни миллионов слов, актуальной становится скорость разметки каждого слова.

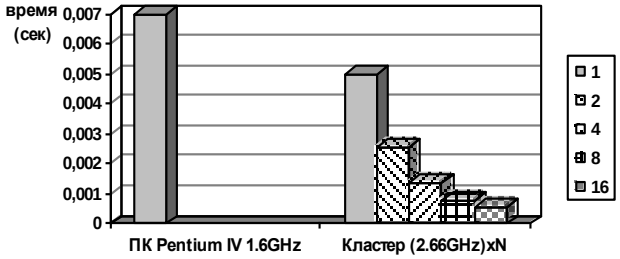
В Институте кибернетики по проекту Академии наук были разработаны кластеры СКИТ-1, СКИТ-2. Они представляют собой вычислительную систему из 16 и 32 компьютеров, называемых узлами кластера [5]. Высокая производительность достигается за счет того, что вычисления разбиваются на независимые блоки (декомпозиция задачи), которые считаются параллельно на разных узлах. Кластеры СКИТ-1 и СКИТ-2 являются открытыми ресурсами для институтов НАНУ, на них решалось множество научных задач, включая экономические, физические, химические. Доступ к кластерам СКИТ осуществляется с помощью Интернета, позволяя использовать его ресурсы находясь на расстоянии от территориального расположения Института кибернетики. Вычислительный ресурс кластеров СКИТ был применен для повышения эффективности автоматической разметки текстов для Национального корпуса украинского языка.

Для того чтобы извлечь пользу от кластеров, необходимо выполнить декомпозицию задачи автоматической разметки – определить, как эта задача сможет решаться

параллельно. Суть решения состоит в том, что входной текст разбивается на части, а каждый узел кластера обрабатывает (размечает) свою часть. После вычислений кусочки размеченного текста «склеиваются» в один размеченный текст. Для организации поиска словоформы в словаре использованы индексы в виде бинарного дерева поиска. Индексы хранятся в оперативной памяти, обеспечивая существенное ускорение обработки.

В результате применения такого алгоритма, была достигнута такая производительность:

Рисунок 1. Зависимость времени разметки одного слова от числа процессоров



Текст, полученный в результате разметки, представляет собой промежуточный материал для создания корпуса. Существуют словоформы, для которых нет соответствия словоформ в словаре (например, имена собственные), не все слова можно автоматически определить (например, сокращения), для некоторых словоформ существует сразу несколько соответствий (слова с одинаковым написанием используются в разном качестве). Сотрудниками Института кибернетики разработано программное обеспечение, позволяющее просматривать и править размеченные тексты, выставлять разметку для слов, разметка которых не существует в словаре, выбирать нужную разметку при наличии нескольких. В результате создается текст в специальной разметке, из который принципиально возможен импорт в любую разметку, которая определяется создателями корпуса.

На данный момент единственным правилом выставления соответствия словоформы ее морфологическому разбору является словарь словоформ. Это позволило корректно разметить 33% текста, взятого в качестве тестовой выборки. 45% слов содержали неоднозначности (для них существовало несколько вариантов разметки). Соответственно 22% слов не были найдены в словаре. Дальнейшим направлением исследований видится нам применением обучающих и статистических методов для формирования правил, используя которые можно уменьшить неоднозначность автоматической разметки.

Литература

1. Демская-Кульчицкая О. М., Семеренко В. Р., Ющенко Р. Р. Методы автоматической разметки текстов национального корпуса языка // Компьютерная математика. – К. – 2004. – С. 70–76.
2. Демська-Кульчицька О. М. Базові поняття корпусної лінгвістики // Українська мова. – 2003. – № 1. – С. 40–46.
3. Brill E. A Simple Rule-Based Part of Speech Tagger //In Proceedings of the DARPA Speech and Natural Language Workshop. – San Mateo, California: Morgan Kauffman, 1997. – P. 112–116.
4. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of speech Tagging. //Computational Linguistics. – 1995. – v.21. – No 4. – P.122–132.
5. Суперкомпьютеры ИК НАН Украины // http://cluster.icyb.kiev.ua.