General and Professional Education
2/2012 pp. 3-12
ISSN 2084-1469

MODES OF AUTOMATING LEXICON COMPILATION AS A COMPONENT OF PRACTICAL STUDIES FOR LINGUISTS

Olga Babina
South-Ural State University,
Department of Linguistics
and Cross-Cultural Communication
454080, Chelaybinsk, prospekt Lenina, 76
e-mail: olga_babina@mail.ru

Abstract: This paper deals with the problem of lexicon compilation in different modes: manual, computer-assisted, automated, automatic ones. The task of completing the morphological zone of a lexical base is considered. Using this task different modes of compiling the lexicon are presented and their appropriateness for the task is estimated basing on the criteria of speed and quality achieved. It is stated that the automated mode fits best the task of lexical-morphological knowledge acquisition. The work for compiling a lexicon in different modes and its assessment is offered as a task for practical studies for prospective linguists.

Keywords: lexical knowledge acquisition, lexicon compilation, automatic mode, automated mode, computer-assisted mode, manual mode, practical studies.

Introduction

Knowledge elicitation is the most laborious and tedious task when building a natural language processing (NLP) system. But it is a necessary part which is a must for getting linguistically appropriate results in NLP-applications. Computational linguists come across this type of work in their professional activity, and this makes topical research tasks of this type for the students of the departments of computational linguistics.

The core of an NLP system is a lexicon which should contain linguistic knowledge sufficient for adequate and robust functioning of the system. The example of such system may be an authoring application for patent claims [Sheremetyeva 2003], machine translation of patent claims [Sheremetyeva 2007], also a system for the assistance in writing abstracts of scientific paper which is being developed by a set of researchers nowadays. In all these application there is a module for generating an output text. For the text to be syntactically and morphologically correct the knowledge base must be flawless with respect of the task of text synthesis. A considerable part of the success of such system is the knowledge base devoid of misprints and inconsistency and the model for text generation that adequately reflects the structure of the language material processed.

When modeling a system the knowledge base design can be biased towards preference of declarative knowledge to procedural one or vice versa. When procedural knowledge is preferred it allows interpolating transformational rules onto the new linguistic material unknown to the system. But the linguistic model may be incomplete or the model may apply wrong rules to some exceptional cases of which the system is 'unaware'. The result is the wrong output. On the other hand, the declarative knowledge which keeps the information about linguistic units in explicit form requires much effort at the design stage but ensures completely correct output for the sublanguage presented in the knowledge base. At the same time, it is restricted only by the knowledge kept in the base. Finding the trade-off between laying emphasis on declarative or procedural knowledge is a key task when developing a lexical knowledge base, including the one for text generation. A future linguist has to learn the difficulties of the development stage and learn to find an adequate solution when developing a lexical base.

Methodological basis for automating lexical knowledge acquisition

A distributed compiling of a lexicon for text generation system is a way to make students see the working process with a real application of computational linguistics and is an excellent learning task for practical studies that prepares them for the work with real projects.

The purpose of any application of computational linguistics is optimizing language function [Баранов 2001], including the recording function which comprises storing lexical knowledge in lexicons. One of the ways to optimize research stage is automating monotonous and formalized procedures for processing language representations:

– automatic compilation of word-lists for lexicons from corpora;
– formulaic expressions extraction;
– automating NP extraction from corpora;
– building a lexical component for a computer-assisted translation system;
– automating the generation of morphological forms for lexical units;
– etc.

The tasks can be divided basing on the performance mode into:

– fully automatic tasks;
– automated task;
– computer-assisted task;
– manual task.

The criterion which allows referring the task to one of these modes implies finding a trade-off for speed and quality that can be achieved by working in one of these modes. Some of the tasks are quite trivial and can be performed fully automatically without a considerable loss in quality, e.g., building a list of word-forms used in the corpus. But most of them require additional linguistic knowledge base which allows solving the problem of automated processing of linguistic evidences. Thus, building a list of lexemes (unlike word-forms) used in the corpus requires the usage of some morphological analyzer (in case of fully automatic task those requiring minimum of pre-requisites for their functioning, e.g., unsupervised learning techniques [Goldsmith 2001], a corpus-based morphological analysis [Бабина, Дюмин 2012], etc.) that could alleviate merging of different morphological paradigmatic forms of the same lexeme (which is most vital for inflectional languages); for formulaic expressions extraction task researchers use different techniques applying criteria of lexical context constraints and term uniqueness [Sheremetyeva 2009], paradigmatic modifiability of terms [Wermter et al. 2005], apply POS taggers and then filter out typical patterns [Kis et al. 2004]; etc. The bottleneck of automatic treatment of non-trivial linguistic tasks is that it provides probabilistic results, with a certain degree of precision which generally does not achieve a 100-percent level. In case when high precision result is required some manual labor has to be resorted to. Then, the performance mode has to be "lowered" from fully automatic to computer-assisted one with pre- or post-processing of the automatically driven results, automated mode with inter-processing of the linguistic knowledge (manual 'interference' into the system functioning in interactive mode), or even fully manual mode in case when correction phase of the knowledge base at any stage of automatic processing is more effort-consuming than when all the task is performed manually from scratch.

All the modes have to be known to practicing linguists, and they must be able to classify the tasks by the mode which is most appropriate for each of them. Junior students of computational linguistics may have initial acquaintance with a ready-made task trying different modes for the same task. As the lexical component of a linguistic knowledge base is the crucial point of an NLP system we will consider the task of building a lexicon in different modes as a task for practical studies of students. The lexicon in focus is a component of a set of NLP applications, e.g., an authoring system for abstracts generation.

The lexicon structure

The knowledge base of the authoring system for abstracts generation includes a corpus-based bilingual (Russian-English) lexicon over a rich feature space, rules and knowledge representation language. The knowledge for the lexicon is being elicited from the corpora restricted to the domain of scientific papers on mathematical modeling. The lexical examples further will be from this domain.

The entry is divided into the parts the number of which complies with the number of languages presented in the lexicon. The entry at each part of the lexicon is described with the

identical set of features each of which, meanwhile, may have different values which is directly dependent on the language. Every monolingual entry is minimally defined with the following features [Sheremetyeva 2007]:

- semantic class – a class that differentiates the meanings of polysemantic words;
- part of speech – includes a label which refers the word to a notional part of speech (noun, verb, adjective or adverb) or a functional part of speech (conjunctions, prepositions, quantifiers, etc.). The set of notional parts of speech is rather standard. The functional parts of speech are determined empirically; the distribution of words among functional parts of speech partially coincides with the traditional classification but for some categories (e.g., wh-words, abbreviation, etc.);
- morphological paradigm – it is a feature which is dependent on the value of the slot Part of Speech as the morphological paradigm is conditioned by lexical-grammatical meaning of lexical units. The lexicon utilizes a whole-word approach for a paradigm representation; all the paradigmatic forms for every lexical unit are explicitly verbalized and stored in the base as indivisible units;
- case-roles (for predicates) – a set of semantic roles relevant for the predicate lexeme;
- fillers (for predicates) – lexical categories that can fill case-roles;
- patterns (for predicates) – a feature that codes co-occurrence of case roles and their linear order in the surface-level syntactic structure.

The set of semantic classes is rather scarce and includes not more than two classes for every notional part of speech (POS) to differentiate the meaning of polysemantic words. It is explained by the constraints of the restricted domain which does not allow the terms to have more than two meanings. Two classes proved sufficient for the restricted domain as two most frequent meanings of words of a notional POS cover almost all occurrences of the word over the corpus.

Morphological paradigm is dependent on the part of speech of the word and the language. Only notional parts of speech (nouns, adjectives and verbs) have a rich paradigm. Forms of the morphological paradigm are presented in the lexicon explicitly for every entry.

Other features are relevant only for predicates. As we are going to concentrate on nouns and adjectives, these features are out of the scope of this paper.

What is non-trivial for the lexicon is that it may contain multi-word linguistic units as a separate entry. The part of speech is assigned to these units depending on the syntactically dominant word of the word combination.

As according to the lexicon structure all the paradigmatic forms for entries of notional parts of speech are stored in explicit form, the main task at the development stage is filling in the morphological zone of an entry with all its paradigmatic forms which is rather time-consuming. Putting aside the features of predicates we shall further present the paradigm of nouns and adjectives in detail as the task offered for practical studies concerns completing the morphological zone of the lexical base for these parts of speech. This subtask will be in focus when practicing different modes of work for lexicon compilation.

The morphological component of the lexicon

We shall consider the task of inputting morphological knowledge into a lexical base. Apparently this stage is preceded by the tasks of acquiring lists of lexemes and deciding which ones should be stored in the lexicon. Most commonly word-lists are acquired from corpora. It may turn out a multi-stage procedure and will require building difference lists to sort out those words already inputted into a lexicon. Moreover, as we take as an initial condition the fact that multi-word units (we concentrate mostly on noun phrases plus some compound prepositions) may function as an entry in the lexicon, it requires application of different techniques to identify appropriate lexical units. This procedure may also be fulfilled in different performance modes (manual, computer-assisted, etc). We shall omit detailed description of preceding stages and focus on the morphological feature of lexical units.

As it has been mentioned before, for inflectional languages the formation of the paradigm for words is not a trivial task. Although there can be observed some analogy in word variation, it is often the case that morphological paradigm of certain words deviates from the common rule, and it is sometimes not evident the factors of what nature are the reason for that. In a working NLP application, the morphological component can

be designed as a built-in component of the system which operates at runtime. But due to the irregularity of some morphological forms and roughness of any model which may not cover all the cases of morphological variation of words, the safe way to deal with morphology is keeping in explicit way all the forms for every entry in the lexicon. This "whole-paradigm" approach is rather time-consuming at the design time but the reward for that is a flawless functioning of the NLP application when dealing with a lexeme known to the system and all its morphological variants.

We shall consider the task of inputting morphological forms for nouns and adjectives in the Russian languages. The Russian language is a synthetic language with a rich paradigm for notional parts of speech (except for adverbs). Nouns in the Russian language have got the categories of number (singular and plural) and case (Nominative, Genitive, Dative, Accusative, Instrumental and Locative). The morphological frame for the noun is formed as a product of the values of two morphological categories (12 forms in total).

Adjectives in the Russian language are described with the features of number, case, animateness-inanimateness (relevant for masculine singular accusative and plural accusative) and gender (masculine, feminine and neuter relevant for singular forms). Some paradigmatic forms coincide for masculine and neuter (all case forms except for nominative and accusative); in the lexicon model these forms were merged. Besides, paradigmatic forms of the adjective in accusative differ depending on whether the adjective syntactically depends on the noun denoting an animate being or an inanimate thing. As a result, morphological paradigm of adjectives is represented by 22 forms (7 case forms for plural, including 2 accusative forms; 2 forms for neuter singular (nominative and accusative), 3 forms for masculine singular (nominative, accusative animate, accusative inanimate), 4 forms for common masculine and neuter singular, 6 forms for feminine singular).

The decision to include multi-word units (noun phrases) into the lexicon puts forward the question of syntactic relations between components of the phrase and corresponding morphological variation of all the components of the phrase. It is decided to include only multi-word lexical units whose components are related with subordination and apposition putting aside the relations of coordination (interdependence is not included as well as it is never present in noun phrases and not relevant for them).

As a result multi-word noun phrases in Russian valid for the compiled lexicon are structured in accordance with the following typical for the Russian language syntactic patterns:

(Pre-attribute)* N_H (Post-attribute)*, where

Pre-attribute = ((Adv)* **Adj**)+ | **Part** ((Prep)* N)* | ((Adv)* Qu)+

Post-attribute = (((Adv)* Adj_{Gen})* N_{Gen})+ | (Prep ((Adv)* Adj_X)* N_X (((Adv)* Adj_{Gen})* N_{Gen})*)+ ,

where: N – a noun (in any case), N_H – the head of the noun phrase, N_{Gen} – a noun in genitive case, Part – a participle, Adv – an adverb, Adj – an adjective, Adj_{Gen} – an adjective in genitive case, Adj_X and N_X – an adjective and noun correspondingly in the same case X, + means that the bracketed chunk can be repeated one or more times, * means that the bracketed chunk can be repeated zero or more times, | makes the border for alternative ways of attribute representation.

For example, the following phrases comply with these constraints: *экспериментальные данные, абсолютно непрерывная функция*, *равноточность измерений, ряд множеств функций*, *приемлемый уровень эффективности, поглощенная доза ионизирующего излучения*, *случайная ошибка в распознавании*, *приведенное в примере число*, *бесконечно много пар*, и т.д.

According to the syntactic rules of the Russian language, the parts of pre-attributes in bold are morphologically dependent on the head of the noun phrase which means they change their morphological form depending on the form of the head. For lexical units that have these attributes not only the head noun changes its form in the morphological paradigm but these dependent units also do. As for other appositive attributes, their constitutive adjectives do agree with nouns, but they are not morphologically dependent on the head. As a result within a certain noun phrase they do not vary and are used in a fixed form. For example, the noun phrase *случайная ошибка в распознавании* is built in accordance with the structural pattern 'Adj N_H Prep N'. The adjective belongs to coordinative attributes and has to change its form depending on the form of the principal noun; the post-positive prepositional phrase is in appositive relation and does not undergo any

changes. The resulting paradigm of the phrase is presented in Table 1.

Table 1. Morphological Paradigm of Nouns

Paradigmatic form	Example
Nominative Singular	*случайн**ая** ошибк**а** в распознавании*
Genitive Singular	*случайн**ой** ошибк**и** в распознавании*
Dative Singular	*случайн**ой** ошибк**е** в распознавании*
Accusative Singular	*случайн**ую** ошибк**у** в распознавании*
Instrumental Singular	*случайн**ой** ошибк**ой** в распознавании*
Locative Singular	*случайн**ой** ошибк**е** в распознавании*
Nominative Plural	*случайн**ые** ошибк**и** в распознавании*
Genitive Plural	*случайн**ых** ошиб**ок** в распознавании*
Dative Plural	*случайн**ым** ошибк**ам** в распознавании*
Accusative Plural	*случайн**ые** ошибк**и** в распознавании*
Instrumental Plural	*случайн**ыми** ошибк**ами** в распознавании*
Locative Plural	*случайн**ых** ошибк**ах** в распознавании*

As one can see the forms of the morphological paradigms for nouns and dependent adjectives are formed by endings variation. Morphological variation of nouns in the Russian language is conditioned by gender and declension of the noun. But genitive plural is independent of declension and is identified empirically. Thus, the following pairs contain nominative singular vs. genitive plural: *сетка – сеток, точка – точек, наука – наук*. All the example words vary according to the same declension scheme. In spite of similar word finalization (including not only the ending proper but the final phoneme /к/ in the preceding morpheme) genitive plural is formed differently which may cause problems for automatic morphological forms generation and will require the development of a more exquisite morphological model taking into account both formal and semantic, etymological and other features.

Adjectives and participles are also changed in accordance with a declension scheme which depends on the ending and the preceding phonetic structure of the morpheme. All the adjectives can be divided into six classes basing on the model they use to form the morphological paradigm (see Table 2).

In these classes adjectives having the ending *–ый* in nominative singular masculine have an unambiguous paradigm which corresponds to class 4. Adjectives ending with *–ой* are divided into two classes depending on the preceding consonant. In case of roots ending with hushes (*-ш-, -щ-, -ч-*) some forms of the paradigm have the vowel *–и–* in their structure (Class 5) according to the spelling rules of the Russian language, for the other adjectives the declension scheme contains the vowel *–ы–* in the corresponding forms (Class 6). Adjectives with the ending *–ий* decline in different ways depending on the phonetic characteristics of the preceding morpheme. Participles and adjectives having the roots or suffixes directly preceding the ending with final hushing sounds correspond to class 1 (*возрастающий, высший, дальнейший*); the most representative group of such attributive words are those with the suffixes *–ащ–* and *–ящ–*. Adjectives having the suffix *-н-* or *-нн-* have the paradigm of class 2 (*внешний, внутренний, последний*). Other adjectives decline as class 3 (*вязкоупругий, ветхий, экономический, мягкий*); the most representative group of Class 3 adjectives are those with the preceding suffixes *–к–* or *–ск–*.

Adjectives when used as a pre-attribute of the head noun in a noun phrase manifest only a part of its paradigm as the forms agree with the noun that does not vary in gender and animateness – these features are permanently assigned to every nominative lexeme over the language.

Performance modes for compiling a lexicon

Let us now consider different possibilities to fill in the slots of the morphological paradigm of nouns (noun phrases) and adjectives which is offered as a task for practical studies of prospective linguists. The morphological frame of the lexicon's morphological component has got a set of slots (12 for nouns and 22 for adjectives) which has to be filled; each is correspondingly labeled.

Table 2. Declension Classes of Adjectives

Paradigm	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
N s m	*задерживающ**ий***	*нижн**ий***	*математическ**ий***	*массов**ый***	*больш**ой***	*развит**ой***
G s c	*задерживающ**его***	*нижн**его***	*математическ**ого***	*массов**ого***	*больш**ого***	*развит**ого***
D s c	*задерживающ**ему***	*нижн**ему***	*математическ**ому***	*массов**ому***	*больш**ому***	*развит**ому***
A s m in	*задерживающ**ий***	*нижн**ий***	*математическ**ий***	*массов**ый***	*больш**ой***	*развит**ой***
A s m an	*задерживающ**его***	*нижн**его***	*математическ**ого***	*массов**ого***	*больш**ого***	*развит**ого***
I s c	*задерживающ**им***	*нижн**им***	*математическ**им***	*массов**ым***	*больш**им***	*развит**ым***
L s c	*задерживающ**ем***	*нижн**ем***	*математическ**ом***	*массов**ом***	*больш**ом***	*развит**ом***
N s n	*задерживающ**ее***	*нижн**ее***	*математическ**ое***	*массов**ое***	*больш**ое***	*развит**ое***
A s n	*задерживающ**ее***	*нижн**ее***	*математическ**ое***	*массов**ое***	*больш**ое***	*развит**ое***
N s f	*задерживающ**ая***	*нижн**яя***	*математическ**ая***	*массов**ая***	*больш**ая***	*развит**ая***
G s f	*задерживающ**ей***	*нижн**ей***	*математическ**ой***	*массов**ой***	*больш**ой***	*развит**ой***
D s f	*задерживающ**ей***	*нижн**ей***	*математическ**ой***	*массов**ой***	*больш**ой***	*развит**ой***
A s f	*задерживающ**ую***	*нижн**юю***	*математическ**ую***	*массов**ую***	*больш**ую***	*развит**ую***
I s f	*задерживающ**ей***	*нижн**ей***	*математическ**ой***	*массов**ой***	*больш**ой***	*развит**ой***
L s f	*задерживающ**ей***	*нижн**ей***	*математическ**ой***	*массов**ой***	*больш**ой***	*развит**ой***
N p	*задерживающ**ие***	*нижн**ие***	*математическ**ие***	*массов**ые***	*больш**ие***	*развит**ые***
G p	*задерживающ**их***	*нижн**их***	*математическ**их***	*массов**ых***	*больш**их***	*развит**ых***
D p	*задерживающ**им***	*нижн**им***	*математическ**им***	*массов**ым***	*больш**им***	*развит**ым***
A p in	*задерживающ**им***	*нижн**им***	*математическ**им***	*массов**ым***	*больш**ие***	*развит**ые***
A p an	*задерживающ**их***	*нижн**их***	*математическ**их***	*массов**ых***	*больш**их***	*развит**ых***
I p	*задерживающ**ими***	*нижн**ими***	*математическ**ими***	*массов**ыми***	*больш**ими***	*развит**ыми***
L p	*задерживающ**их***	*нижн**их***	*математическ**их***	*массов**ых***	*больш**их***	*развит**ых***

* N – nominative case, G – genitive case, D – dative case, A – accusative case, I – instrumental case, L – locative case, m – masculine gender, n – neuter gender, f – feminine gender, c – the form which is common for masculine and neuter gender, s – singular number, p – plural number, an – animate, in - inanimate.

Manual mode

Fully manual mode for completing the slots of the morphological frame implies typing in all the forms one-by-one in each slot. This mode promises manual control over the inputted entries, which requires maximum of human participation at the compilation phase. However, full control does not guarantee a flawless knowledge base due to the human factor (fatigue, loss of attentiveness in the course of time, mechanical misprints, etc.) which leads to the rise of inconsistency in the base. Moreover, the absence of any automation makes this mode ineffective in terms of time resource needed. At this mode the higher quality of the compiled lexical base free of misprints and inconsistency can be achieved by further sacrificing time performance.

Computer-assisted mode

In the Russian language (and many European languages as well) morphological paradigmatic forms of the word are made with the help of endings rather regularly changing for different words of the same part of speech. The stem remains unchanged which prompts copying the fixed parts of the lexical units fully automatically into all morphological slots at a time. On the automatic phase being completed a correction or addition of endings is left for the linguist. As endings do not have one-to-one correspondence to the part of speech, normally the words have to be further subdivided into declension classes, the procedure for completing stems to fully-fledged word-forms may be performed manually or with the help of the designed morphological model which employs a morphological classification in supervised or unsupervised mode.

In the computer-assisted mode the endings are being post-edited manually. For example, for the lexical unit *линейн-ый эллиптическ-ий оператор-* the unchanged stems of its components are copied into 12 slots of the morphological paradigm of a noun and then every stem is completed with the corresponding endings for the slot: *-ый –ий –Ø* for nominative singular, *-ого –ого –а* for genitive singular, etc.

Due to the automatic phase for copying unchanged stems this mode considerably benefits in time when compared to manual mode. Moreover, it is less prone to misprints as

the linguist has to concentrate on correct spelling for the majority of the linguistic material once for each entry and not when typing every form of the morphological paradigm. The slots completion is again under a full manual control which allows minimizing incorrect form generation, but still human factor remains although to a far less degree. Misprints when applying this method appear practically only in the endings of word-forms.

Automated mode

The automated mode is similar to the computer-assisted mode in the sense that the operation is fulfilled in the combined way. Unlike the computer-assisted mode the automatic phase and manual correction phase are inter-switched, and the task is performed with human participation at the initial, middle or final stages of editing the entries of the lexicon.

We enhance the automatic phase with the help of empirical morphological generator. The procedure of entering an entry into the lexicon in automated mode consists in: 1) the user enters the unit in one of its paradigmatic forms (normally in Nominative singular); 2) for each separate word of the new entry: a) the morphological generator enlists candidate model words that, probably, have the same declension model as the inputted entry; in case no form is found the generator offers the default model word as a candidate; b) the user has to choose the correct model word (the word which indeed has the same declension scheme as the inputted entry); c) the morphological generator automatically fills in the slots of the morphological paradigm of the inputted word according to the declension scheme of the model word chosen by the user at the previous step. In case there is no model word with the correct declension scheme, the user may form the paradigm manually and add the current word into the database as a model word. Some components in multi-word units may remain unchanged for every form of the morphological paradigm, e.g., in the noun phrase *неточность в распознавании* only the first word is declined while the included prepositional phrase *в распознавании* keeps unchanged. In this case the user may simply add these components to every paradigmatic form in the initial typed-in form.

When enlisting the candidate model words, the morphological features for the model word are also shown to the user as in each part of speech there are ambiguous forms the endings of which may coincide while they belong to different slots of the morphological paradigm, e.g. Adjective Nominative masculine singular *большой* – Adjective Instrumental feminine singular *абстрактной* – Noun Genitive feminine singular *кривой*. The morphological features shown are restricted to the feature of case as it has been empirically determined that the case feature is sufficient to identify the word and its form.

The declarative knowledge of the morphological generator used in computer-assisted mode contains model single-word units with their paradigmatic forms. The procedural knowledge contains two types of rules: a) rules according to which the generator filters out candidates from the declarative knowledge base for the user to choose which declension model the current entry follows; b) rules according to which it generates paradigmatic forms of the entry word similar to the paradigm of a model word from the declarative knowledge base. Both types of the knowledge base are language-dependent.

At the initial stage the declarative knowledge base is empty or contains only a paradigm of a word with the default declension scheme. The declension scheme is represented by the endings that the word has in different paradigmatic forms. This knowledge is represented implicitly with the help of a model word in the declarative knowledge base having the forms corresponding to a certain declension scheme and the rules in the procedural knowledge base that allow splitting the model word into a stem and the endings of its paradigmatic forms and transfer this model onto the paradigmatic frame of a newly inputted entry.

The declarative knowledge base is being collected simultaneously with the process of the lexicon compilation. As adjectives and nouns have got different sets of features in their morphological paradigm, knowledge bases are collected separately for a) single-word adjectives and b) nouns and their component (including adjectives that are restricted by a certain gender and feature of animateness-inanimateness prescribed according to the noun they are dependent on).

For the Russian language the initial base of adjectives is empty; default declension scheme of nouns is represented by the words of masculine gender with the zero-ending in

nominative singular (such as *алгоритм, вектор*, etc). The user inputs other declension schemes while filling in the paradigms when compiling the lexicon. In case of a multi-word unit each component of the unit should be changed to form its morphological paradigm. As soon as the forms of the entry are completed, it may be saved in the database as a model. Next time when the user enters a word with the same morphological paradigm as the model word saved in the database this model word may be extracted with the help of the rules in the procedural knowledge base and offered to the user.

According to the filtering rule of the procedural knowledge base, when the user enters a new word the generator compares, first, the last three and, next, the last two characters of the current word with the endings in the forms present in the declarative knowledge base. In case a form with the same ending is found in the base the rule adds it to the list of candidate model words for the current entry.

After the user chooses one word from the candidate list, the morphological generator identifies the longest identical ending (two or three characters) for the model word and the word from the current entry. Accordingly, it splits the form of the model word into the stem and the ending, finds the same ending in the word-form of the current entry and considers the rest of the current word-form as a stem. The current word-form is assigned with the morphological features corresponding to the model word-form chosen by the user. By splitting the identified stem from every word-form of the model word, adding the remained as a result endings to the stem of the current entry word and putting the resulting forms into the relevant slots, the morphological paradigm for the current word is built.

Thus, when working in this mode human participation consists in inputting a lexical unit in one of its paradigmatic forms, choosing for every component of the unit the model word from the base having the same declension model and updating the database as soon as a new declension scheme is detected. The computer assists by automating the process of generating morphological forms for every component of the inputted lexical unit and filtering out from the database the probable words with the same declension scheme as the current entry.

This mode retains all the benefits of the computer-assisted mode. It helps to almost completely avoid spelling mistakes. Taking into account that not only stems are automatically copied but the endings are also generated automatically, on the one hand, the problem of misprints in manually typed endings is solved; on the other hand, the time is saved to a greater degree. Manual control over the chosen declension scheme ensures choosing the correct variation model.

Automatic mode

Fully automatic mode comprises inputting the words and their generated morphological paradigm in unsupervised way. It is the most time-consuming at the development stage as it requires analyzing a representative corpus of linguistic examples to model the morphology of the language and design the morphological generator that realizes this model.

The automatic mode is the most desired in terms of efforts saving at runtime stage. It allows compiling a full morphological lexicon for the predefined class of lexical units within seconds. Actually, compiling a lexicon with the usage of the module of fully automatic morphological paradigm generation is rather senseless, as just as well this module can be used at runtime stage. Then it will not be necessary to keep in the lexicon full paradigm of the lexical unit, which saves the space.

The bottleneck of this mode is the quality. For a lexicon-based system the lexicon is the core, the quality of which greatly influences the performance of the system. Processing in automatic mode is prone to mistakes as the language is full of exception cases, and any formal model usually operates within some restricting conditions. Human control cannot be completely excluded, otherwise quality suffers.

Evaluation of the Modes Appropriateness

Manual and computer-assisted mode in the form presented above are rather slow and inefficient ways to solve the task of compiling a lexicon that comprises full morphological paradigm of the entries.

The usage of the automated mode with inter-editing is preferred to automatic one for noun phrases. Due to the existence of ‘irregular’ forms of genitive plural the number of declension scheme for nouns rises compared to distinguished four declensions in the traditional

grammar. Thus, the words *сетка – точка – наука – тройка* that would traditionally be referred to the same declension belong to four different declension schemes, as one of their forms is made differently for each of these words, even though the endings for all the other paradigmatic forms coincide. As for genitive plural there are at least four cases for the variation of words with this declension – the inclusion of the unstable vowel *–о–* before the final consonant of the stem and zero ending (*сеток*); the inclusion of the unstable vowel *–е–* before the final consonant of the stem and zero ending (*точек*); zero ending (*наук*); the interchange of the consonant *–й–* into the vowel *–е–* before the final consonant of the stem and zero ending (*троек*). This problem, to be solved automatically, requires a rather complicated morphological model which may not cover all the cases possibly existing in the language.

Moreover, there are cases of ambiguity for different paradigmatic forms, e.g., Noun Nominative plural *множества* – Noun Genitive masculine singular animate *Сидорова.* To identify which form it is in each case the system requires human assistance. In automatic mode this problem can be solved only by prompting the user to input all the words in the same paradigmatic (basic) form, for example, nominative singular. But it puts forward the problem with the restricted paradigm when the word has only plural (e.g., *данные, часы*) or only singular form (e.g., *уклончивость*). The sets of paradigmatic slots valid for these two classes never overlap, which makes any choice of the basic form inappropriate for one class of words or the other.

The decision to input multiword units which may contain nouns, adjectives and other parts of speech also sets the problem of inter-POS ambiguity, e.g., the pairs of words Noun *края* – Adjective *логическая*, Noun *уравнений* – Adjective *верхний*, Noun *семейство* – Short Adjective *хаусдорфово*, Adverb *медленно* – Noun *волокно* should belong to the same declension scheme according to the rule of equality of two last characters, which is not true. Some of these cases are impossible to solve on the basis of the form only. The syntactic context could be helpful but again it requires a new model to process similar words and at some stage corpora are difficult to access; what is processed are ready-made word-lists.

It seems that the automated mode helps to achieve the highest possible rate for noun phrases without sacrificing the quality of the acquired knowledge.

Adjectives when acquired separately, not within a multi-word noun phrase, are easier to identify. Full adjectives can be divided into six classes (see above), which are quite unambiguously detected basing on the form characteristics of the morphemes preceding the ending. For research purpose we use the morphological generator of Russian adjectives that utilizes as the declarative knowledge base this classification. But, again, depending on the stress which is normally not marked graphically in the text, some adjectives may have in the form of nominative masculine singular the ending *–ий* in the unstressed position or the ending *–ой* in the stressed position, which influences the forms variation. In automatic mode this ambiguity problem could be solved by restricting the adjective input by the basic form only. In case the adjectives are extracted from the corpus where they can be used in any of their paradigmatic forms, it requires pre-editing the list of inputted adjectives. Another way is, taking into account that in the corpus analyzed there was no adjectives with the ending *–ой* in masculine singular, to exclude class 5 and 6 from the processing algorithm. Apparently, the absence of such adjectives in the previously analyzed language material does not guarantee the same in the future, which is why this approach requires post-editing phase at which all the automatically generated entries are checked and corrected if necessary. Thus, to minimize effort and ensure high output quality adjectives should be inputted into the lexicon in automated mode with pre-, inter- or post-editing of automatically generated adjectival paradigmatic forms.

Conclusion

The task to fill in the slots of the morphological component of a lexicon in four modes described represents a possible way to organize practical studies for prospective linguists.

As the ratio speed-to-quality is one of the crucial points for the development of modern linguistic software the evaluation phase of different modes should also be included into the task. The most evident parameters that can be evaluated for each mode are: a) how fast the task is fulfilled; b) how often mistakes appear

c) what problems there are when performing the task in this mode.

The first question implies timing. To answer the second question one of the additional elements of the practical studies that may be added is cross-checking the partner's results. The third question is open-ended and may reveal different problems, including both linguistic material and the procedures to process it.

Generally, the evaluation of different modes shows that the most reasonable way to perform the task is the automated mode which allows saving time by avoiding mechanical typing in the paradigmatic forms. Simultaneously, human control over the compiling procedure ensures high quality of the compiled lexical base.

References

1.Goldsmith, John. Unsupervised Learning of the Morphology of a Natural Language. In *Computational Linguistics*. Volume 27, Number 2. 2001. Pp. 153-198.

2.Kis, Balazs, Begoña Villada, Gosse Bouma, Gábor Ugray, Tamás Bíró, Gábor Pohl, John Nerbonne. A New Approach to the Corpus-based Statistical Investigation of Hungarian Multi-word Lexemes. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004. Lisbon, Portugal. 2004. Pp. 1677-1680.

3Sheremetyeva, S. Towards Designing Natural Language Interfaces. In *Proceedings of the 4th International Conference "Computational Linguistics and Intelligent Text Processing"* Mexico City, Mexico, February 16-22. 2003.

4.Sheremetyeva, S. On Probability of Resources for a Quick Ramp up of Multilingual MT of Patent Claims. In *Proceedings of the MT Summit XI Workshop on Patent Translation*. Copenhagen, Denmark. 2007. Pp. 28-33.

5.Sheremetyeva, S. On extracting multiword NP terminology for MT. In *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation* / ed. Lluís Màrquez and Harold Somers. Universitat Politècnica de Catalunya, Barcelona, Spain, May 14-15. 2009. Pp. 205–212.

6.Wermter, Joachim, Udo Hahn. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver, Canada. 2005. Pp. 843–850.

7.Бабина, О. И., Н. Ю. Дюмин. Корпусный метод автоматического морфологического анализа флективных языков // Вестник ЮУрГУ. Сер. Лингвистика. – 2012. – Вып. 14. – С. 4-9.

8.Баранов, А.Н. Введение в прикладную лингвистику. М., 2001.