Wim Honselaar and Marijke Elstrodt, University of Amsterdam

The electronic conversion of a dictionary: from Dutch-Russian to Russian-Dutch

ABSTRACT: The paper consists of two parts. The first part deals with the output of the conversion and its lexicographic peculiarities. In the second part we discuss some technical aspects, including the computer programs for the extensive analysis of the original entries and for the actual conversion.

Introduction

Since 1989 the University of Amsterdam has been working on the compilation of a large Russian-Dutch dictionary (= R-D dict.) which will become the counterpart of Van den Baar's large Dutch-Russian dictionary (= D-R dict.). As this latter dictionary was available in computer readable form, we decided to "convert" the D-R dict. electronically. The product of this conversion should be a rough preliminary version of the R-D dict., which would, of course, require further modification and amplification.

The printed version of the D-R dict. was produced on the basis of a text file. As an example, the entry UITHANGEN (hang out) is shown in Figure 1 in the typography of the D-R dict., and, only partially, in text format in Figure 2.

uithangen *I tr. (vlag e.d.)* вы́весить$^{201(A)}$; *(kledingstuk, lakens e.d.)* широко́ разве́сить^{201A}; *(z. voordoen als)* разыгра́ть (из себя́) [4]; ◆ *de idioot —, (ook)* вали́ть^{1} *if.* дурака́; *II intr. (van vlag e.d.)* висе́ть^{162Б} *if.*; ◆ *de vlag hing uit, (ook)* был вы́вешен флаг; *overal hing(en) de vlag(gen) uit,* везде́ бы́ли разве́шены фла́ги; *waar hangt hij uit?, (persoon, voorwerp)* *куда́ он де́лся / дева́лся?; *где он торчи́т?; *куда́ он пропа́л?; *где он запропасти́лся?.

Figure 1

The entry consists of 2 sub-entries, for transitive and intransitive UITHANGEN respectively. Transitive UITHANGEN has several translations, all of which are provided with collocational or semantic information in order to make clear the differences in meaning and/or use. In this sub-entry there is only one phraseologically bound translation. Intransitive UITHANGEN has one translation but several phraseological units. The superscript codes in the entry refer to the appendix of the dictionary which contains examples of more than 400 inflectional and derivational paradigms. The asterisk is used as a marker of style and stands for "colloquial".

<headw>	uithangen	(hang out)
<syntax>	I tr.	(transitive)
<semant>	vlag e.d.	(flag, ...)
<transl>	vyvesit'	
<morpho>	201(A)	
<semant>	kledingstuk, lakens e.d. (clothes, sheets, ...)	
<transl>	shirokó razvésit'	
<morpho>	201A	
<semant>	zich voordoen als	(pretend to be)
<transl>	razygrát'	
<morpho>	3	
<cont.tr>	(iz sebjá) (4)	
<phrase>	de idioot —	(play the fool)
<sem.phr>	ook	(also)
<tr.phr>	valját'	
<morpho>	1	
<grammar>	if.	(imperfective aspect)
<cont.tr>	duraká	
<syntax>	II intr.	(intransitive)
<semant>	van vlag e.d.	(of flag, ...)
....		
<phrase>	waar hangt hij uit?	(where is he hanging out?)
<sem.phr>	persoon, voorwerp	(person, object)
<style>	*	(colloquial)
<tr.phr>	kudá on délsja / deválsja?	
<style>	*	(colloquial)
.....		
<end>		

Figure 2

In Figure 2 every line represents a field, which is identified by a label, such as HEADW(ORD), MORPHO(LOGY), TRANSL(ATION) etc., each field containing a specific part of the text of the entry. There are separate fields for morphological, semantic/collocational, syntactic and stylistic information: MORPHO, SEMANT, SYNTAX and STYLE respectively. The fact that specific information is stored in specific fields made it

feasiable to convert the dictionary electronically. This conversion did not, unfortunately, involve a simple reordering of the fields, but required an extensive analysis of the structure of all entries of the D-R dict. The technical aspects of the conversion will be dealt with in the second part of this paper. In the first part we will discuss the product of the conversion. We will be dealing with three questions: What does this product look like? Is this type of conversion efficient? Does conversion give a better product?

1. Lexicographic aspects of the conversion

1.1. The product

The output of the conversion consists of an enormous number of what we call MINI-ENTRIES, see Figure 3. A mini-entry consists of a Russian headword, morphological codes, semantic, syntactic and stylistic notes, and, last but not least, a Dutch translation. When a headword occurs more than once, a polysemy number has been added. Phraseological units have been (semi-)electronically supplied with a relevant headword. Finally, all mini-entries have been alphabetically ordered.

What is remarkable in the mini-entries is the presence of detailed semantic notes with many Russian headwords and phraseological units. These notes contain synonymous equivalents for the translations or collocational restrictions of the kind of "with respect to ...".

Bilingual dictionaries with the mother tongue of the users as the target language only rarely contain such detailed information; the translations are supposed to be sufficient for understanding and translating the foreign language. Even in the case where a word has more than one distinct meaning, and, consequently, two or more translations, semantic information is not always supplied. For many users, however, especially advanced students and translators, this kind of additional explanation will give more insight in the meaning of Russian words and in the selection restrictions of these words. An example is "persoon, voorwerp" (person, object) in field <SEM.PHR>, specifying that the phrase "kudá on propál?" in the entry PROPÁST' may refer to both animate and inanimate subjects (see Figure 3).

In some cases the notes may later be changed to alternative translations. A good example of this is the note "zich voordoen" (pretend to be) in the entry RAZYGRÁT' which is a translational equivalent to "uithangen" (although the style of "zich voordoen als" is more formal than the colloquial "uithangen").

It is clear that the converted version will need extensive editing. Firstly, a lot of material will have to be deleted such as the numerous "explanatory translations" of words for which the editors were unable to find a straight Russian equivalent. Clearly, many of these have a specifically Dutch cultural source. Secondly, many words and phraseological units will have to be added, especially Soviet terminology and vocabulary, words relating to the Orthodox church and native culture. Thirdly, many more or less obsolete words from the great Russian literature of the 19th century will have to be added, as well as many new words which appeared during the last few years. Finally, we will have to include the "normal" words, units and meanings which for no specific reason at all happen to be left out of the D-R dict.

<headw>	visét'	
<morpho>	162B	
<grammar>	if.	(imperfective aspect)
<syntax>	intr.	(intransitive)
<semant>	van vlag e.d.	(of flag, ...)
<transl>	uithangen	(hang out)
<headw>	vyvesit'	
<polysem>	1	
<morpho>	201(A)	
<syntax>	tr.	(transitive)
<semant>	vlag e.d.	(flag, ...)
<transl>	uithangen	(hang out)
<headw>	vyvesit'	
<polysem>	2	
<phrase>	byl vyveshen flag	
<tr.phr>	de vlag hing uit	(the flag is hanging out)
<headw>	dét'sq	
<phrase>	kudá on délsja / deválsja?	
<style>	*	(colloquial)
<sem.phr>	persoon, voorwerp	(person, object)
<tr.phr>	waar hangt hij uit?	(where is he hanging out?)
<headw>	durák	
<phrase>	valját' (if.) duraká	(imperfective aspect)
<tr.phr>	de idioot uithangen	(play the fool)
.....		
<headw>	propást'	
<phrase>	kudá on propál?	
<style>	*	(colloquial)
<sem.phr>	persoon, voorwerp	(person, object)
<tr.phr>	waar hangt hij uit?	(where is he hanging out?)
.....		
<headw>	razygrát'	
<morpho>	3	
<phrase>	razygrát' (iz sebjá) (4)	
<semant>	zich voordoen als	(pretend to be)
<transl>	uithangen	(play/act)
.....		

Figure 3

1.2. Efficiency

The electronic conversion of the dictionary required a lot of preparatory work. The investment of approximately one man- year is, however, cost-effective, because the electronic conversion yields an enormous saving in typewriting (keyboarding), proofreading and correction, as well as an even greater saving in cumbersome research into the semantic, syntactic, morphological and stylistic characterization of Russian words and phraseological units, and in finding corresponding Dutch translations.

1.3. Better product

Apart from the savings mentioned before, the conversion has been very profitable in other respects. We will confine ourselves to two areas. The first area concerns the set of headwords in the R-D dict. The D-R dict. contains many Russian expressions from contemporary informal language, an area that is poorly reflected in standard Russian dictionaries. Because of this, certain areas of the Russian language are inaccessible to foreigners. Thanks to our conversion, however, many colloquial expressions will appear in the R-D dict. The second area concerns the use of the dictionary as a translator's dictionary. In many cases a Dutch word was translated into Rusian as a more or less ad-hoc collocation. For these Russian ad-hoc collocations, the converted dictionary contains specific Dutch expressions (words, collocations or idioms) which are not normally used for the translation of Russian collocations or idioms, and which are, as a consequence, now "available" to be used in a Dutch translation of Russian. An example of this is the translation of "oná ogryznúlas' v otvét" into "zij reageerde met een snauw" (lit. she reacted with a snap: she snapped at someone). Usually, OGRYZNÚT'SJA is translated as "afbekken" (snap at), see Van den Baar (1979); the occasional combination with "v otvét" (in reply) leads to the translation "met een snauw reageren".

Conversion provides other sources of information for the translator. An example is when the equivalent of a Dutch singular word is a Russian plural. In the converted dictionary, the plural will be included separately in the entry. A Russian translation for the Dutch word SUPPORTER (supporter) is "bolél'shchik" and the Dutch singular collective noun AANHANG (supporters) is translated as "bolél'shchiki", the plural form of "bolél'shchik". The R-D dict. will present the possible translation of the plural form as:

bolél'shchik	supporter; (in plur. ook) aanhang. (supporter; (in plur. also) supporters)

2. Technical aspects of the conversion

2.1. The choice of the computer programs

Because no existing computer program was available for the conversion of the D-R dict., new tools had to be developed. The properties of the designs were deduced by analyzing the descriptions, the proposed set of transforming operations and the content of the computer files of the dictionary. This revealed that: 1. there was an almost endless variety in the specific composition of the entries; the structure of a substantial part seemed to be

even more complex than the informal specifications suggested; and 2. most of the proposed operations were highly context sensitive. Moreover, the order in which the operations had to be applied seemed to be a crucial one; every change in the order produced a significant restatement of the content of the composing operations.

It became obvious that apart from a conversion program, we also needed a tool for the analysis of the structure and the details of the composing entries. Both programs had to be developed simultaneously; implementation of the transforming operations would raise questions about structural details and a new understanding of the structure would have consequences for the content and the ordering of the proposed transformations.

One option was to develop computer programs written in a well-known programming language like "C" or "Pascal". Such programs, however, lack flexibility. Every change of the original design requires an extensive adaptation and a long series of test runs and debugging sessions, especially when recursive and/or ordered actions are involved. Such a large investment in time would not be justifiable for programs that are to be used only once.

The most obvious choice, therefore, seemed to be the use of the program system Parspat, developed at the Computer Department during the last 10 years. This system, designed for a variety of tasks, has two main areas of application:

1. formal description of existing texts
2. transformation of texts, including substitution, deletion, multiplication and/or reordering of the composing parts.

In the past few years, we gained a lot of experience by using it as a tool in different areas of research. Among other uses, it was succesfully applied in the development of a grammar for the English language and in a program for transforming a written Dutch text into composing phonemes.

The Parspat system is a program generator, consisting of a compiler and a run time system. The programs that can be written in this system have a rather simple syntax. Usually the source text is rather small, parts of it can be tested separately and changes are relatively easy to perform. The programs are, in fact, formal grammars: a set of interdependent rewriting rules, that define and describe patterns in the input. Brainteasers like recursive actions and ordering problems are solved to a great extent by the system itself. Instead of a technical description, which can be found in Van der Steen (1988), we present in the next two paragraphs some examples extracted from our programs.

2.2. Exploration of the structure

The Parspat system enables the step by step construction of a formal description of a large amount of text. In a top-down approach, starting from a global scheme, one can gradually explore details.

As an example we consider a part of the informal description of the dictionary files:

- each file consists of a number of entries, seperated by an end of lemma indication.
- each entry consists at least of one Dutch word (sometimes supplemented by notes), followed by a one or more Russian translations (each with possible notes); some entries contain two or more of such series.
- some entries conclude with examples of Dutch phrases with their Russian translations.

This (simplified) description was implemented in a Parspat program in the following way[1]:

```
(1) File      :: (Entry)+ .
(2) Entry     :: (DutchPart , (RusPart)+)+ , (Example)* .
(3) DutchPart :: DutchField , (DutchNote)* .
(4) RusPart   :: RussianField , (RussianNote)* .
```

At the same time, by attacking the problem in a bottom-up manner, one can combine the separate entities in the input (tokens, words, sentences) into larger ones. For example, the fact that a morphological code consists of a number, sometimes followed by a letter, can be translated to:

```
(5) Morpho    :: ( 0..9 )+ , ( A..Z ) .
```

Each version of the descriptive program was run on the computer with the dictionary files as input. In most cases, the outcome "Deviations found" was followed by restatement and/or refinement of the description. However, it turned out that the original files, although constructed by means of an extended editor, contained many inconsistencies and typing errors. We also observed redundancy in the information stored in the entries. In order to adjust every entry to the over-all structure, some transforming Parspat programs were applied by which the original files were repaired and compressed. In a relatively low number of cases the imperfect fields had to be improved manually. As a major result, the transforming operations were adjusted to cover the deviations from the original description.

2.3. The conversion programs

At this stage we had at our disposal an empirically tested detailed description of all entries, reasonably "clean" and compact files ready for the conversion and a newly written and extended series of operations.

The actual conversion was performed by a cascade of transforming Parspat programs, linked together to one performing program. The first program transforms the original files, and each consequent program acts on the output of the preceding one.

Like descriptive programs, a transformation program consists simply of a set of rules. Every transformation rule has two parts: an input side (on the left-hand side of the transformation operator ">") and an output side (on the right-hand side). The input side defines a pattern in the same way as in the descriptive rules: only parts of the input that match this description will be changed into the pattern described at the output side of the rule. The input can be the original material, the output of a preceding transforming program or the output of an already activated and performed rule in the same program. The units of which each side of a rule is composed may be tokens, words, fields and even groups of fields. The meaning of each unit must be stated somewhere in the program using a descriptive rule.

As an example, we take a look at a simplified version of one of the most crucial operations. In the converted dictionary, the original order between the Dutch word and its Russian translation had to be reversed. This is done by a rule like:

```
(6)     DutchPart , RusPart , LemmaEnd >
        RusPart , DutchPart , LemmaEnd .
```

Of course, descriptive rules that define what is meant by "DutchPart" and "RusPart" must be added to the same program. In fact, those defining rules were extracted from the descriptive programs, discussed in the preceding paragraph.

As an illustration of the easy way by which a transforming Parspat program can be written, we consider the implementation of a recursive action. The power of such a rule and its clarity is based on probably the most important principle of the system: a rule is performed ONLY and, at the same time, AS LONG AS an input matches the input conditions of that rule. In other words, if a rule has been applied to some input and the (transformed) output again matches the input conditions of the same rule, the operation is simply repeated. As soon as this process ends, the final output is ready for treatment by other rules.

With rule 6, the order of the Dutch and Russian parts can only be reversed in a small number of special entries: those in which the Dutch word has only one translation. Most entries, however, contain two or more Russian translations. Each of them must be transformed into a separate entry and the Dutch word must be multiplied. In the case of exactly two translations this could be accomplished by:

```
(7a)    DutchPart , RusPart1 , RusPart2 , LemmaEnd >
        RusPart1 , DutchPart , LemmaEnd ,
        RusPart2 , DutchPart , LemmaEnd
```

Instead of stating a separate rule for entries with 1, 2,..n translations, one may cover all possibilities by using recursion. This rule 7b is an adapted version of 7a:

```
(7b)    DutchPart , RusPart , RestLemma , LemmaEnd >
        RusPart , DutchPart , LemmaEnd ,
        DutchPart , RestLemma , LemmaEnd
```

Crucial is the defining rule for the entity "RestLemma", stating that it may consist of one or more Russian parts:

```
(8)     RestLemma :: ( RusPart )+ .
```

The effect of rule 7b is that only the first translation becomes a separate entry; the rest of the original entry remains unchanged for the moment. As long as the rest contains two or more translations the same rule will be applied over and over again. This process will stop as soon as the remaining part contains only one translation, because at that moment the input side of rule 7b does not match the input any more. At that moment rule 6 becomes applicable; by that rule only the order of the Dutch and Russian parts will be reversed.

3. Conclusion

The conversion of the D-R dict. required a relatively small investment in time but has resulted in a very useful new product of about the same size as the original dictionary. The approach of converting the D-R dict. by using Parspat programs has been a success; rather complicated descriptions for transferring a variety of different entries into a new structure were easily implemented. Indispensable for this process was the detailed screening of the original files by a series of descriptive programs.

Endnotes

1 The meaning of the meta symbols is:
:: – "is defined as"
> – "to be transformed into"
, – "followed by"
[..]+ – "one or more times the enclosed entity"
[..]* – "zero or more times the enclosed entity"

Bibliography

VAN DEN BAAR, A. H. (1989): Nederlands-Russisch Woordenboek. Kluwer, Deventer-Antwerpen. (1758 pp.)

VAN DEN BAAR, A. H. (1979): Russisch woordenboek, Russisch-Nederlands. Coutinho, Muiderberg. (502 pp.)

VAN DER STEEN, G.J. (1988): A program generator for recognition, parsing and transduction with syntactic patterns. CWI tract 55, Centre for Mathematics and Computer Science, Amsterdam.

KEYWORDS: computational lexicography, bilingual dictionary, program generator, syntactic pattern matching, active text filters.