Lexin - a report from a recycling lexicographic project in the North
Ann-Kristin Hult, Sven-Göran Malmgren and Emma Sköldberg
Centre for Lexicology and Lexicography, Department of Swedish, University of Gothenburg

In the late 70s, the Swedish Board of Education initiated a project (the Lexin project) aiming at production of dictionaries between Swedish and many immigrant languages. A monolingual Swedish dictionary was compiled, serving as the common base of the bilingual dictionaries. In the 90s, the project was exported to other Nordic countries. Since the Nordic languages are closely related, much of the work carried out in Sweden could be reused in Norway, Denmark, and Iceland. Today, there are many learners’ dictionaries between Nordic languages and ‘exotic’ immigrant languages, especially with Swedish and Norwegian as source languages.

In this paper, we account for some aspects of this - in some respects probably unique - project. At the end, we give a description of the revision and updating of the Swedish database that has been going on since 2008.

1. Introduction

Some 30 years ago, the Swedish Board of Education initiated a project with the object of producing dictionaries particularly suited for the teaching of Swedish to immigrants. Its basic motivation was the increasing need for learners’ dictionaries and bilingual dictionaries among the many immigrants coming to Sweden from non-Scandinavian countries. The project, led by Dr. Martin Gellerstam, was named Språklexikon för invandrare (Dictionaries for immigrants, Lexin for short). One important goal of the project was to compile a well-designed Swedish vocabulary provided with various kinds of linguistic information. The Swedish vocabulary would serve as a basis for a series of bilingual dictionaries. In turn, these bilingual versions would allow for a reversal of the vocabulary, not only allowing Swedish but also each foreign language to be the source language of the bilingual dictionaries. The latter goal, however, was not fully realized (Gellerstam 1999:3-4; cf. below).

The Swedish Lexin material then also became a matter of Nordic interest. In the early 1990s, the Swedish database was made available to all the other Nordic countries in order for them to start producing their own national Lexin dictionaries. It was assumed that if you could find equivalents of the Swedish Lexin material in the immigrant languages, then it should also be possible to change source languages, enabling re-use of the translations already made into the target languages in question. Considering the close relationship between Swedish, Norwegian, Danish, and to a certain extent Icelandic, it was believed that a substantial part of the Swedish material could be re-used. This would save both time and money, which is essential when compiling dictionaries of immigrant languages with relatively few speakers living in Scandinavia (cf. Holmboe 1999; Bjørneset & Svaravsdóttir 2003:52-53, 63).

Today, the initially Swedish Lexin project has developed into an interesting Nordic project of lexicographic cooperation. Since the beginning of 2007, the Swedish Language Council is responsible for the comprehensive Swedish part of the project. By now, the Swedish part of the project includes no less than one monolingual and 15 bilingual printed and/or electronic dictionaries (see below).

In 2008, the Language Council initiated a project involving updating and improving the Swedish dictionaries. First of all, a revision of the monolingual Swedish database had to be

---

1 Thanks to Christian Mattsson for providing us with valuable information on the Swedish Lexin dictionaries. We also want to thank Tove Bjørneset and Ásta Svaravsdóttir for providing information concerning the Norwegian and Icelandic parts of the Lexin project.
made. This work is carried out at the Centre for Lexicology and Lexicography at the University of Gothenburg. Based on the new version of the Swedish database, existing bilingual dictionaries will be revised, and new ones will be compiled.

Unfortunately, the Lexin projects have received little attention outside the Nordic lexicographic environment. The purpose of this article is to describe in more detail the basic idea behind the Lexin project and how it has been put into practice in the Nordic countries. We focus on the work carried out in Sweden, where the project started. We will, above all, discuss the original version as well as the new version of the Swedish dictionary and database.2

2. The Swedish Lexin series

The Lexin dictionaries are primarily aimed at immigrants who have recently come to Sweden and are about to learn Swedish. The dictionaries are intended to be easy to use even for persons with limited reading abilities and little or no experience of dictionaries. Consequently, in comparison with many other dictionaries, a user-friendly layout is especially important. (See Beijer & Fre Woldu 1997:4; Gellerstam 1999:5.)

2.1. The monolingual Swedish database and the paper dictionary Svenska ord

The Swedish database was compiled at the Department of Computational Linguistics, University of Gothenburg. From the database, a paper dictionary (Svenska ord ‘Swedish words’) was generated in 1984. Later editions appeared in 1992 and 1995. From now on, when talking about Svenska ord, we refer mainly to the paper dictionary, but sometimes also to the database. The second and third editions contain approximately 28,000 lemmas, the lemma selection being adapted to the needs of the dictionary’s target group in a very thorough way. It includes, for example, common words with related derivatives and frequent words from textbooks and teaching aids. All lemmas have been provided with comprehensive information typical of a general dictionary. An entry will be quoted below.

In addition, Svenska ord includes specific words covering special issues and institutions in Swedish society, along with factual information, for example LO (‘the Swedish Trade Union Confederation’), Högsta domstolen (‘the Supreme Court’) and the names of the political parties. The dictionary is mainly intended for reception, but also for production; the former purpose is above all served by definitions and examples, the latter mainly by information on pronunciation and construction patterns (cf. Gellerstam 1999:5-6). About 1,700 elementary words are described by means of pictures in a special section. This is made up of 31 themes, for example Family (and relatives), The human body (exterior and interior), Mammal and insects, and also Cooking and meals. These themes are available in a separate edition, too. They are, of course, extremely important for users with very poor knowledge of Swedish, but they can also serve as pedagogical support for more advanced learners or even Swedish-speaking users with reading and writing difficulties (cf. Bjørneset 2001, Pálfi & Tarp 2009).

Late in 2009, a paper on learners’ dictionaries in Scandinavia appeared in Lexicographica (Pálfi & Tarp 2009). After approvingly describing the Lexin project, the authors propose a number of improvements, some of which had actually already been implemented in the new version of the Swedish database. We will discuss some of Pálfi & Tarp’s proposals at the end of this paper.

---

2 Late in 2009, a paper on learners’ dictionaries in Scandinavia appeared in Lexicographica (Pálfi & Tarp 2009). After approvingly describing the Lexin project, the authors propose a number of improvements, some of which had actually already been implemented in the new version of the Swedish database. We will discuss some of Pálfi & Tarp’s proposals at the end of this paper.
Since 1994, there has been a free online version of the Lexin series (http://lexin2.nada.kth.se/). The online version ‘talks’; a human voice gives the pronunciation of all the lemmas. Moreover, besides the pictures there are animations classified into fifteen main sections. In this way, verbs otherwise difficult to make comprehensible through illustrations - let alone definitions - are made much easier to understand. By clicking on Kitchen and cooking, for example, film sequences show the semantic distinctions between the Swedish equivalents of verbs like cut, chop, mash and grate, which can otherwise be hard to describe through an ordinary picture.

From the start, the online version has been used to an ever-increasing extent. Today, the site has about one million searches/month (http://lexin2.nada.kth.se/statistik.html; cf. Kann 2004). In a log-file based user study of the online Svenska ord dictionary, Hult (2008b) points to the dictionary’s limitations as far as more advanced searches are concerned. She also accounts for the large number of unsuccessful searches that users are faced with in the online version. Many of these search strings are correctly spelled Swedish words, but they are not included in the limited lemma list. Further, the user is not always linked to the entry when searching for inflected forms of the lemma. The results indicate the necessity of an update of the dictionary’s contents as well as the online version’s search options (see also Dahl 1999:47).

In Hult (2008a) the log-file analysis of Svenska ord is followed up by a web questionnaire addressed to the users of the dictionary (about 350 questionnaires were sent in). The users were asked, firstly, basic questions concerning age, gender, native language, etc., and, secondly, questions about their searches in the dictionary. The majority of the users were students between 15 and 29 years of age. The informants represented 49 different languages. However, considering the fact that Lexin’s target group is made up of immigrants, a surprisingly large proportion of the informants (60 %) had Swedish as their mother tongue. Many of the users probably do not view Svenska ord as a dictionary for immigrants but rather as a user-friendly free online dictionary. The study also shows that the information types most requested are definition and spelling, followed by examples.

2.2. The bilingual dictionaries
As mentioned earlier, the monolingual Swedish database has provided the basis for Lexin’s bilingual dictionaries. The idea of using the same source language material as a basis for many bilingual dictionaries is by no means uncontroversial, but it is not unique to the Lexin project. An ambitious Dutch project has produced a number of bilingual learners’ dictionaries with Dutch as source or target language, based on a common monolingual Dutch database. It is especially interesting that several dictionaries with Dutch as target language have been semi-automatically reversed from the original dictionaries with Dutch as source language. (See Martin 2007 for a discussion of the so-called CLVV project, and Maks 2007 for a presentation of the OMBI editor.)

Today (February 2010), there are 15 bilingual dictionaries available via Lexin’s website (see Table 1; dictionaries written in italics are being planned but are not yet available).
The selection of target languages is, among other things, related to the number of immigrants and the existence/non-existence of conventional dictionaries. The reason for compiling a special Swedish-English dictionary was that English, for many students, functions as a mediating language between their mother tongue and Swedish (cf. Beijer & Fre Woldu (1997:5); cf. also Martin (2007:222-226), presenting the selection principles of the CLVV project.)

In compiling a bilingual Lexin dictionary, several problems have to be addressed. For instance, the language in question may have different varieties or orthographic systems, and it is not always easy to choose the ‘right’ one. A good example is the Swedish-Spanish dictionary. Most Spanish-speaking immigrants in Sweden come from Latin America, so it was obvious that a Latin-American variety of Spanish should be chosen. In view of the great number of Chilean refugees from the 70s onwards, the dictionary was, to some extent, based on their variety of Spanish. It must be remembered, however, that the production time for a Lexin dictionary is several years, and that the streams of refugees can change very rapidly. Thus, you cannot always be certain that the variety chosen satisfies the real needs when the dictionary is completed. (Cf. Beijer & Fre Woldu 1997:4, 11-12, 24-25).

The need for dictionaries of new languages has constantly grown. When financial resources have been limited, smaller dictionaries have been compiled, including only 5-8,000 lemmas, e.g. the Swedish-Vietnamese dictionary. The bilingual dictionaries contain the same information categories as the Swedish database, irrespective of size. It has not been a trivial
task, however, to reduce a dictionary of medium size (i.e. *Svenska ord*) to a small dictionary (cf. Malmgren 1999).

As is the case with the Swedish online version, the bilingual dictionaries have been used frequently, especially the Swedish-English dictionary, but also the Swedish-Arabic, the Swedish-Persian, and the Swedish-Spanish dictionary.

As mentioned above, several dictionaries with Dutch as source language have been successfully reversed within the CLVV project. Unfortunately, no such reversal has been carried out within the Lexin project. One reason is that it has been necessary to prioritize the production of Lexin dictionaries with new immigrant languages as target languages, partly due to new streams of refugees (See Gellerstam 1999:11). It should be emphasized, however, that the Internet versions of the bilingual dictionaries with Swedish as source language can be used as dictionaries with Swedish as target language. For instance, a Turkish user can search for a Turkish word among the Turkish equivalents in the Swedish-Turkish dictionary, and get the Swedish translation (See Pálfi & Tarp 2009).

3. Lexin in the other Nordic countries

Below, we briefly present the status of the Lexin project in other Nordic countries, and also comment on experiences concerning the transfer of the Swedish database to other Nordic languages.

3.1. Lexin in Norway

Norwegian work on Lexin started in 1996 (Hovdenak 2008:219). In Norway, there are two monolingual Lexin dictionaries, one in Standard Norwegian and the other in New Norwegian. Both varieties are available on Lexin online. Nine bilingual e-dictionaries have been compiled with Standard Norwegian as source language and ten e-dictionaries with New Norwegian as source language. Two examples are the Standard Norwegian-Somali and the New Norwegian-Tigrinya dictionaries. Apart from the online versions, there are four printed bilingual Lexin dictionaries. However, future dictionaries will be published only as e-versions, for example a Norwegian-Vietnamese dictionary ([http://decentius.hit.uib.no/lexin.html](http://decentius.hit.uib.no/lexin.html)). Hovdenak (2008:219) points out that the Norwegian Lexin dictionaries are not widely known, but that the use of the online version has increased substantially in the last few years.

Bjørneset & Svavarsdóttir (2005:53ff) report that more than 80% of the database of Norwegian could be translated manually from the Swedish database. However, Bjørneset (2002:36) emphasizes that Norwegian social and administrative institutions are often not directly comparable with the Swedish ones:

> A number of terms and words relating to Swedish society (institutions, social structure, culture, etc.) have been inactivated in the Norwegian database. To compensate for this, records for comparable and related Norwegian concepts have been added to the database. (Bjørneset 2002:36)

Also, many new words have been added to the Norwegian database. As a consequence, it currently has about 8,000 more lemmas than the Swedish one3. There are also, as Bjørneset (2002:40) points out, some structural differences between these two languages which had to be taken into account in the transfer process. Finally, in the Norwegian project a set of new

---

pictures has been created, since some of the Swedish pictures were too culture-specific. (For further discussion, see Bjørneset 2005.)

Bjørneset & Svavarsdóttir (2005:63) also report that, in compiling the Norwegian-English dictionary, 80 % of the Norwegian database could be linked to the existing Swedish-English translations. The Norwegian-Tamil dictionary, by contrast, had to be translated manually, which of course made this dictionary considerably more expensive to produce.

3.2. Lexin in Iceland

In Iceland, the Lexin work started with a pilot project in 1998. In spite of the fact that Icelandic differs from Swedish to a much larger extent than does Norwegian, it turned out that more than 60 % of the Swedish lemmas could be translated to corresponding Icelandic words (Svavarsdóttir 1999:22). The pilot project also showed that the words’ formal properties naturally needed to be adapted to the rich and varied set of inflectional patterns in Icelandic (Svavarsdóttir 1999:24ff). Since 2001, the Lexin projects in Norway and Iceland have worked closely together. For example, the Icelandic project is working directly in the Norwegian data base application (Bjørneset & Svavarsdóttir 2005:53ff.).

3.3. Lexin in Denmark

The Danish Lexin project is considerably smaller than the projects in Sweden and Norway. It includes four bilingual e-dictionaries (see http://lexin.emu.dk/): Danish-Swedish, Danish-Turkish, Swedish-Danish and Turkish-Danish. Thus, the Danish project has adopted the idea of vocabulary reversal and offers dictionaries with Danish as target language. However, these dictionaries comprise only about 7,000 lemmas. It seems that to a large extent the Danish project has benefitted from the work on the pictures carried out in Sweden. In other respects, however, it is difficult to comment on the Danish experiences, since they have been poorly documented. According to Holmboe (1999:61), the Danish project has had the same experiences and arrived at the same solutions as the Norwegian project.

4. On the updating of the Swedish Lexin database

Let us now take a look at an entry in Svenska ord (1992) and the corresponding entry after the ongoing revision. We choose the entry väntar ‘wait(s)’. The older version of the entry is followed by an English translation.

![Table: Entry väntar with translation](image)

The first thing to be noticed is the form of the verb lemma; not the infinitive, but the present tense. (In the English translation, the 3rd person present form has been chosen as an equivalent.) As Gellerstam (1999:7) points out, this is due to the fact that immigrant teachers traditionally have recommended the present as the ‘remember form’. In fact, it is easier to
generate the other inflectional forms from the present form than from the infinitive form. For instance, if you know the present form *väntar* (see the example), you can create all the other forms by means of grammatical rules. But if you know the infinitive form (*vänta*), you cannot be sure that the present form is *väntar* (rather than, e.g. *vänter*).

As representatives of the inflectional paradigm, four forms have been chosen in the old version of *Svenska ord*, namely the past, supine, infinitive, and imperative. The last two forms, in this case, coincide in Swedish, and have been merged (*wait(!)* means, of course ‘to wait’ or ‘wait!’).

There are two numbered senses of the lemma *väntar*. The definition of the first one runs something like: ‘let time pass till something happens’. It is followed by valence information (A, B = ‘somebody’, x = ‘something’, S = clause) and examples, the first two self-explanatory syntactic examples, the third an idiom (with explanation). The fourth example is a derivative, the deverbal noun meaning ‘the act of waiting’.

Let us now turn to the revised version. (Only relevant details are translated - in the text.)

<table>
<thead>
<tr>
<th>väntar</th>
<th>[Pven: tar] verb, att vänta, väntade, har väntat, är väntad, vänta!</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>låter tiden gå tills något inträffar, avvaktar:</td>
</tr>
<tr>
<td>-</td>
<td>de fick vänta på bussen i tio minuter</td>
</tr>
<tr>
<td>-</td>
<td>han väntade på besked från sin läkare</td>
</tr>
<tr>
<td>&lt;ngn väntar (på) ngn/ngt&gt;</td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>väntar barn är gravid</td>
</tr>
<tr>
<td>2</td>
<td>tror att något ska hända:</td>
</tr>
<tr>
<td>-</td>
<td>han väntade inte att han skulle komma in på högskolan</td>
</tr>
<tr>
<td>-</td>
<td>kostnaderna väntas öka nästa år</td>
</tr>
<tr>
<td>&lt;ngn väntar ngt/att+S&gt;</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2. Entry väntar revised version**

In the new version of the entry, several changes have been made. First, inflectional information has been made somewhat more explicit; the infinitive and imperative forms are no longer merged, and a new form, the past participle, has been added. Also, in order to facilitate understanding of the information, markers have been inserted before the infinitive, the supine, and the past participle forms, respectively (*att* ‘to’, *har* ‘has’, and *är* ‘is’). In Swedish, contrary to English, the difference between the supine and the past participle is overt, it being important to learn this difference.

Further, the definition is now followed by a synonym (*avvakta* ‘wait’). Generally, a great many synonyms and antonyms have been added in the new version. This is mainly in order to increase users’ chances of understanding the entry words; one single synonym - even if not a perfect one - can make the difference between understanding and not understanding. Besides, the synonyms and antonyms also serve the purpose of increasing users’ vocabulary.

Also, the form of the main verb of the definition has been changed from infinitive to present tense (*låta > låter* (‘lets’)). The purpose is to make definitions more compatible with the lemma forms.

After the definition, the structure of the article has been somewhat changed. In the new version, compounds - not represented in this entry - come first, followed by self-explanatory syntactic examples and then idioms. Idioms are now given in bold, emphasizing their status as
a kind of sub-lemmas. The deverbal noun väntan has been transformed into a full lemma, since it is not self-explanatory. Valence information - now somewhat more transparent - is to be found at the end of each numbered sense in the new version.

In a learner’s dictionary mainly used in order to understand what words mean, i.e. for reception purposes, the role of the examples is extremely important. Not infrequently, it is impossible to write a simple definition, comprehensible to users with a limited vocabulary. In these cases, good examples are indispensable in leading the user in the right direction. Normally, the examples should be full sentences, rather than infinitival phrases. They must be self-explanatory, not containing difficult words (and if possible, not long compounds, even if the elements of the compounds are familiar). Ideally, they should evoke a little scene from everyday life, with prototypical actants (more or less in accordance with FrameNet principles).

It is very important that the difference between production and reception dictionaries is clear. To take, for the case of simplicity, the English word crime, the example commit a crime is of absolutely no use as long as the dictionary is consulted for reception purposes. Of course, the collocation is helpful to users who want to write a text containing the word crime, but before reaching that level, they must understand the word. A ‘full’ example like he committed a crime and had to go to jail would be much more helpful.

Most examples in the older editions of Lexin are good. Nevertheless, several infinitival phrases have been replaced by full sentences, e.g. in the entry väntar. For example, instead of the infinitival phrase vänta på bussen (‘wait for the bus’), the full sentence de fick vänta på bussen i tio minuter (‘they had to wait for the bus for ten minutes’) is given. To the user who does not quite understand the definition, it should be helpful that the sentence contains the three main actants of WAITING: a person who waits, something (s)he is waiting for, and a certain amount of time.

Interestingly, we have, independently, often arrived at the same conclusions as Pálfi & Tarp (2009). For instance, they propose more synonyms and antonyms, more explicit valence information, lemmatization of idioms (we promote idioms to sub-lemmas), and the use of a controlled vocabulary in definitions. True, we do not make use of a defining vocabulary in the Longman sense, but we try very hard to avoid difficult (and long) words when defining.

On one point, however, we slightly disagree with Pálfi & Tarp. They recommend numerous examples related to valence information, quoting from Swedish corpora. In principle, this is a good idea (although very time-consuming), but their concrete examples are, in our view, far too lengthy and difficult. For instance, as an example of the verb pratar ‘talks’, they propose ‘Johnny bowlar med sin kamrat och pratar med honom om att han träffat en tjej på babymatiken’. (‘Johnny is bowling with his friend and talks to him about the fact that he has met a girl at the baby rhythms’.)

Finally, search options of the Internet versions will be considerably improved. For instance, every entry will be easily accessible from any inflectional form of the corresponding lemma.

5. Conclusion: advantages and disadvantages of Lexin

The Swedish Lexin project has generated one monolingual dictionary and a large number of bilingual dictionaries; it has also given rise to interesting collaborative projects within Nordic
lexicography. Evidently, the project’s basic idea - to develop a Swedish vocabulary base for bilingual dictionaries - is not only applicable within the Swedish project but also possible to transfer to the other Nordic countries. The concept has been labour-saving and has created economic opportunities for the production of bilingual dictionaries with relatively few potential buyers. As Hult (2008a, 2008b) shows, the Swedish vocabulary base has functioned as an independent dictionary for many L2 as well as L1 learners of Swedish. However, the user studies carried out by Hult demonstrate the material’s limitations and also that the search engine interface can be further improved.

Over the years, certain drawbacks of the Lexin project have been discussed. For example, one might have principled objections to its fundamental idea. It might be argued that, ideally, the Swedish lemma selection in each bilingual dictionary should be based on a contrastive analysis of the target language and Swedish (Malmgren 1999:79). Problems have also arisen with some of the translators involved, due to their limited lexicographic experience. As a result, the Swedish bilingual dictionaries are of varying quality (Gellerstam 1999:13).

The great similarities between the Nordic languages have been an important precondition for this Nordic lexicographic collaboration. Nonetheless, as Bjørneset & Sívavarsdóttir (2005:54) point out, the differences between the languages are greater than might be thought. One thing is clear, however: the work that has been accomplished, above all in Sweden, has benefitted all the Nordic countries involved. Lastly, other (groups of) countries may draw on the Nordic experiences in this field.
References

Kann, V. 2004. ‘Folkets användning av Lexin - en resurs.’ 2010-02-25

Web addresses

http://lexin2.nada.kth.se/ [Access date: 26 Febr. 2010]
http://lexin.emu.dk/ [Access date: 26 Febr. 2010]
http://lexin2.nada.kth.se/statistik.html [Access date: 26 Febr. 2010]
http://decentius.hit.uib.no/lexin.html [Access date: 26 Febr. 2010]

Note. Emma Sköldberg’s work on this paper was carried out as part of her position financed by the Swedish Academy.