Drug terminology

A multilingual term database. The AVENTINUS project.

Christian Sjögreen
Språkdata/Dept. Of Swedish language
Göteborg University
Box 200
SE-405 30 Göteborg
Sweden

Abstract

This paper starts with a brief overall presentation of the AVENTINUS project, merely a list of the different included modules and some comments. Then follows a discussion about drug terminology and finally a description of the design and implementation of a tailored multilingual drug terminology database in MS Access. The used tags and links are presented and discussed and the inputting situation is described. Then some numerical details of the database and a short concluding remark are given.

Project description

The AVENTINUS Project[1] aims at supporting an Advanced Information System for Multinational Drug Enforcement. It is funded by the European Union in the Linguistic Engineering (LE) Program, and has several development and user partners. The goal of the project is to support drug enforcement with multilingual linguistic expertise. AVENTINUS will support communication by providing linguistic tools to overcome language communication barriers. Users should be able to access information and receive results of search requests in their own native language, even if the information is derived from foreign language sources.

The languages dealt with in the first phase are English, German, Spanish, and Swedish.

AVENTINUS will provide modules and components that can be linked to and integrated into the users domestic environments. Modularity and integratability are the most prominent features of the software solutions to be provided.

The participating users are domestic police organisations and intelligence agencies and the Europol Drug Unit (EDU). Interest from other authorities have also been noticed, though.

[1] For an exhaustive description of the project, see [THUR97] or take a look at our AVENTINUS web page at

The data to be handled by AVENTINUS are of several types, the most important ones for AVENTINUS being textual data. Texts from **open sources**, mainly newswire texts and **internal communication** texts, like police reports, will be considered.

According to the User Requirements Report, there are two main scenarios that should be supported, the **Indexing (Data Entry) Scenario** and the **Retrieval (Analysis) Scenario.**

In the Indexing Scenario, users are confronted with incoming texts from different sources (fax, telex, electronic) in different languages. They have to decide whether a given input text is relevant or not. AVENTINUS will support this scenario by providing translation support tools, and by providing information-understanding tools (indexing, information extraction).

In the Retrieval Scenario, search support will comprise tools like name search (transliteration, similarity of names) or text search both in structured and textual databases. Translation support tools will be responsible to translate the search requests as well as the search results (structured or textual) into the native language of the searcher

To support the scenarios, AVENTINUS will provide different types of components, such as **Translation Support, Information Processing Support,** and **Search Support.**

The project will have three types of translation support tools, *Term Substitution, Translation Memory,* and *Machine Translation.* All of them will be available as stand-alone tools accessing common lexical resources, and as components to be called from standard Windows editors such as WinWord. There will be several components to process the incoming texts, and provide further information for later retrieval, for instance *Information Extraction* and *Indexing*

Search Support refers to several requirements, for instance (i) *requests in natural language,* as well as in some *structured form,* (ii) *requests in a native language* instead of the foreign languages of the database to be searched and (iii) *query expansion* and navigation possibilities in the area of text search. Search in both structured databases and in a textual ones will be supported. The components to be offered comprise the following: *Name Search, Search in texts,* and *Search in structured databases.* In order to support the AVENTINUS application, three types of linguistic resources will be set up which have to do with both multilingual issues and domain modelling: *Lexical Database, Thesaurus,* and *Domain Model*

The architecture of AVENTINUS follows two basic principles. It must be based on components that can be integrated in a very flexible way into the existing system environments of the users and it must be very flexible in the interaction of the internal components. In many components the AVENTINUS functionality may be called from a standard text processing system. The interface will be available on several platforms.

A first version of the AVENTINUS data pool, including test texts and terminology, is implemented, as a pool to create resources and test specifications. The complete system specifications

have been written and reviewed, some of the AVENTINUS components (translation memory, machine translation) are operational and a test plan is available, with Europol as the first testing environment.

Drug terminology

Our assignment in the project is currently twofold. One is to collect, structure, link and 'linguistically' edit the drug terms for all languages involved. The other one is to develop linguistic resources for Swedish. Only the former will be discussed in this paper. It is mainly the responsibility of Maja Lindfors Viklund, Yvonne Cederholm and myself, where my job is and has been to design, implement and maintain a database to store and link the terms locally.

Drug terminology [MLV97] differs from 'normal' terminology in a substantial way–as probably most criminal terminology does–as it's partly used not to make communication easier but rather to hide facts. The fact that many of the terms are slang words or argot only emphasises this difference still more. Normally, one can assume a terminological environment to cover a rather specific, well defined domain, and to be rather consistent with respect to ambiguity and stability in meaning and also often in growth. In the case of drug related terminology we face a quite different situation. The domain includes such opposite areas as street slang, police and custom vocabulary, drug legislation, medical treatment, complex chemical compounds. New products–based upon new chemical formulas (referred to as *designer drugs*)–are constantly developed to keep the trade ahead of the legislation since a drug is not prohibited in our society until it's explicitly put on the list of illegal drugs, i.e. classified as narcotic.

The terms in our drug terminology cover areas such as:

- *cultivation* – geographical areas, traditions and methods
- *production* – handling of substances
- *substance* – types of drugs, names
- *trade* – related places and persons
- *equipment* – tools and accessories
- *abuser* – often nicknames according to the drug in question
- *symptom* – behaviour and experiences

We have until today collected some 13,000 terms all together. Mostly English (roughly 5000), and Swedish (4000), but also some 2000 German and about the same amount of Spanish. Since we work in Sweden it has apparently been easier for us to collect Swedish terms than terms from the other languages involved and that explains why we, relatively spoken, have rather many Swedish ones. It is also quite natural that the number of English terms is equally high, or in fact higher, since we are using English as a pivot language. In addition there is the obvious influence from the Anglo-American cultures upon the western European countries, reflected both in life and language. The relatively low amount of German and Spanish terms is, to some extent, explained by the fact that we still haven't come that far in collecting terms from these languages.

When the project was designed, it was assumed that the users, mainly police and intelligence organisations and the EDU, should provide us with texts and lists of terms. But in reality this has not worked very smoothly, perhaps due to the rather delicate nature of these texts and lists. Most of the police organisations have been very reluctant to give anything away. This is understandable but problematic. There are some exceptions though. The Swedish police organisations have been very co-operative and have supplied the project with a lot of material. For instance, we have a very good relationship with Svenska Narkotikapolisföreningen (SNPF), with the Swedish officials in Europol, and with the Swedish National Police Academy. They have supplied us with a lot of materials and in return, I might add, we have had some opportunities to help them [Hol97].

Among other things SNPF has given us–on diskette–the text to their book *Basfakta om narkotika* [SNPF96] (Basic facts about narcotics). The book deals with almost everything in this context and we have been able to extract numerous terms for drugs, tools, treatment of addicts, legislation, etc. from this text. Another main source for the collection of drug related terms is *Internet*. It's amazing what you can find there. Price lists, recipes, articles, etc. etc. So, surfing the Internet has become more work than pleasure for us, at least in this respect. We have also collected newspaper articles about drugs and other related topics and set up word lists and concordances to find more terms.

The 'tagset'

The initial bulk of terms that we collected were prepared in a unix environment and imported into the database where the terms then have been further analysed. The tags we use are:

1. *Part of speech*
2. *Type of term* – D: 'drug term', G: 'General language term'
3. *Language* – EN, ES, DE or SE
4. *Definition* – given in English and/or in a 'native' language
5. *Subdomain* – type of drug
6. *Concepts* – links to the domain model
7. *Comment* – free comment
8. *Original language* – Chinese, Swahili, Inca, ... whatever (not used for the moment)

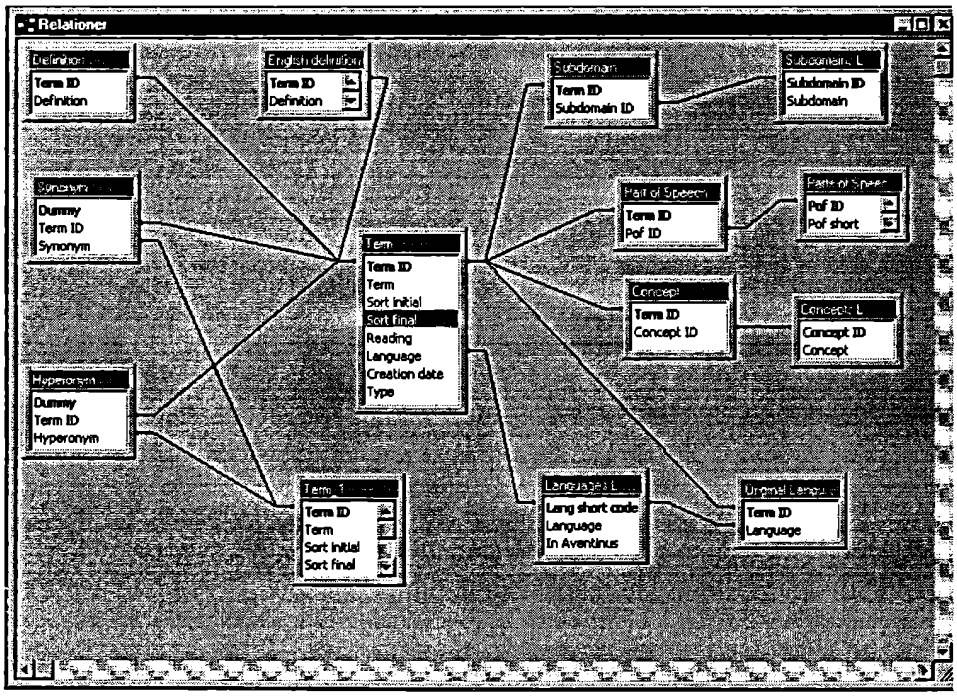
The three first tags are compulsory. The rest are, from the database point of view, optional. Of course, when a term is fully analysed, all relevant tags should be given. The *subdomain* tag is used to classify the terms into drug categories like *cannabis*, *cocaine*, *heroin*, *opium* etc. The *concept* tags classify the terms into categories within the area of drugs and related (criminal) activity and behaviour, for instance *drugs*, *tools*, different types of *geographical locations*, *organisations*, *persons* etc. The set of concepts is decided upon in co-operation with the AVENTINUS group at the university in Sheffield who works with the *ontology* (i.e. the world model, or rather in this case the domain model). The concept tags are hence the links from the drug terminology to the ontology in the project.

We use three semantic links, of which the first one is optional in the same sense as above, while the two others are truly optional.

- Language *equivalence* – to a term in the pivot language
- *Synonym* – to term(s) in the same language
- *Hyperonym* – to term(s) in the same language

The database

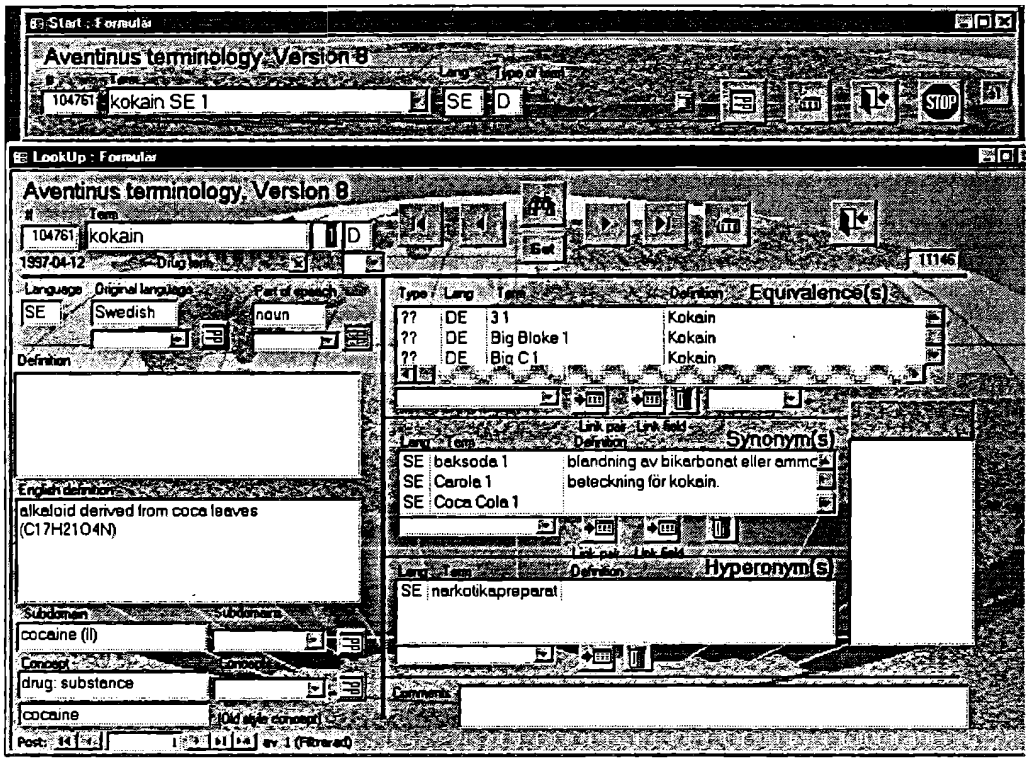
The terms and the tags are stored in a relational database implemented in MS Access 8.0. The main table is called *Term* and contains a unique identification number, the term itself, a normalised form, mainly used for sorting purposes, a reading number (in case of homography or polysemy) and a short form for language. The concepts and subdomains are stored in separate tables *Subdomains L* and *Concepts L* with unique identification numbers and the links in special link tables *Subdomains* and *Concepts* with pointers to *Term* and *Subdomains L* and *Term* and *Concepts L* respectively.



Database relations

There are other tables in the database than those shown in the figure. For instance the rather important table *NextAvailableNumber*, that provides new terms, and some other entities as well, with unique ID-number. All links between terms and properties use the id-numbers instead of the terms themselves. This technique ensures that all established links are kept unchanged even if the terms are corrected for misspellings or if the ordering of reading numbers are changed or whatever. To ensure that no links are set unexpectedly these ID-numbers are never allowed to be reused. That is, if a term for some reason is deleted from the table *Term*, the freed id-number is just thrown away.

The interface



The AVENTINUS Drug Enforcement Input Screen

Existing terms are selected in the smaller upper window in the list field, i.e. a field that lists all entries in the table *Term* in alphabetical order. By keying in one or more letters in the field, the list positions itself to that part of the alphabet. When a term is selected, all tags are automatically fetched and made available for editing in the bigger window.

To enter a new term one clicks on the *New term*-button in either of the two windows. Then a separate input window appears that forces the user to input not just the term, but also the *type* and the *part of speech*. This window will also give information about the presence of any other homographs to the one given and if so gives a warning.

When a non-English term is entered it shall be linked to an English equivalent term. If none exist we try to find one somewhere in our sources. If that isn't possible we can in most cases get a link through a hyperonym and as a last resource we invent a term by 'plain translation'. The linking is carried out by selecting the target(s) from a list field that contains the already existing terms in the same language (synonyms and hyperonyms) or in English (equivalents).

When an equivalence link is established, all other terms in the other two non-pivot languages that are linked to the same pivot term will automatically pop up in the *Equivalence* field. Thus one can get a picture of the popularity of a term (= drug) in the other two non-pivot languages. In the example of the Swedish term *kokain* in the figure there are no less then 88 links, of which one is the pivot term (in English) while the other represent the bouquet of synonyms for this popular drug in German (66) and Spanish (21). The corresponding list of (15) terms in Swedish is shown in the synonym field. To see the English synonyms we just pick the English counterpart as the main entry. In the case of *cocaine* there are no less than 99 synonyms.

When a non-English term is linked to a pivot term in English, it will automatically be linked to a *subdomain* and a *concept* via the already established links from the pivot term. Since not all these terms are linked yet–the project is still in progress–there is an optional possibility to link 'manually' as well.

To each term a definition may be given. The definitions may be given in English and/or in a native (normally Swedish) language. We got a set of terms from the German BundesKriminalAmt (BKA) and some of them were given a kind of rough definition in German. The 'native language definitions' are merely used as a way to give a preliminary definition, perhaps to be translated later or even handed over to a (human) translator. The main purpose of the definitions are today to serve as an aid and a bases for classification and linking.

At irregular occasions, the whole database is exported into textfiles and transmitted via the 'net' to GMS in Munich, where the projects main database is implemented.

Statistics and examples

Here are some numbers, drawn from the database. Note that these numbers may differ from what is mentioned in the text, depending on the fact that all terms are not yet included into the database and certainly not fully analysed. Statistics like these reflect the presumed fact that the amount of terms for a certain drug, or class of drugs, is related to the usage in a society. Perhaps we will eventually find something more exciting than this rather obvious fact. But it's too early to go into this yet, the data has to be more complete first.

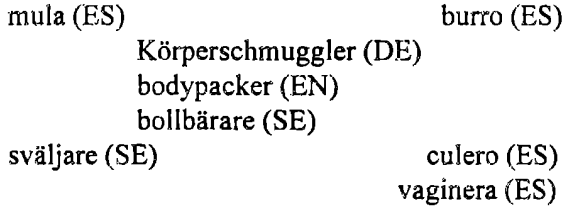
		English	German	Spanish	Swedish	Total
	Terms	5414	2181	1719	1829	11143
	Linked to equivalence		1860	985	1421	4266
Subdomains:	Amphetamine	283	35	3	49	370
	Cannabis	871	68	29	74	1042
	Cocaine	600	55	20	56	731
	Heroin	247	73	15	35	370
	Opium	95	59	4	22	180
Concepts:	Drug: substance	2631	545	110	229	3515
	Drug: tool	65	3	1	18	87
	Person: dealer	55	11	4	24	94
	Person: user	108	7	11	43	169

The following examples from the database are thoroughly discussed in [MLV97].

Inheritance example.

English	German	Spanish	Swedish
deal	deal	dilear	dila
flip (out)	ausflippen	flip	flippa
snort	snorten	snortar	snorta
speed (n)	Speed	espid	speed

Smuggler example.



The 'smuggle' terms all means in principal the same thing, a person who smuggles drugs, but vary from the 'pack animals' via some rather neutral terms to the extremes that describe how, and even where in the body, the drugs are smuggled. This can be seen as an illustration of how cultural attitudes may be reflected in language.

Capital letter example. H is used as short for heroin[2] and has then been expanded to *horse*, perhaps to indicate the strength in the drug. The same term then shows up in the other languages as *H* and *Pferd*, *H* and *caballero*, and *H* and *häst*.

Conclusions

The area of drug terminology is to our knowledge a rather unexplored field of research and shows some interesting deviations from 'normal' terminology. All the linking of terms to synonyms and hyperonyms and to English 'equivalents' results in a network of relations between terms that in many respects are not yet fully taken advantage of. So, when the database is 'ready', I think we will have a good framework to start some rather interesting issues.

[2] The figure *8* (*H* being the 8:th letter of the alphabet) is also used as a synonym for *heroine*.

There have also been suggestions to broaden the database both with respect to other languages and to introduce new domains. The database could then be used as a foundation to create an international police thesaurus. This may give us a lot to do in the future.

It seems to us that the Anglo-American influence on the drug vocabulary is more pronounced in newly introduced drugs, whilst older more established drugs tend to develop a domestic vocabulary. If this is a tendency that will survive is however hard to tell.

The work has been rather tedious and often quite difficult since none of us are a multilingual drug abuser with knowledge in slang from the streets of Hamburg, Liverpool, Barcelona or Mölndal. We have got good help though, from policemen and from people we know, with insight into the domain and/or the different languages.

At last I want to emphasise the use of MS Access in a number of different areas of research. It can be used to store small or medium sized databases and can easily be learned to master by linguists with a little help from their friends. It has become my standard desktop database tool.

Citations

[Thu97] Thurmair, Gr.: *Multilingual Information Processing*: The AVENTINUS system. Paper given at the FBI conference in Berlin, Sep. 1997.

[MLV97] Lindfors Viklund, M.: *Drug terms*. Dept. of Swedish language, Göteborg University 1977.

[Hol97] Holmén, S.: *Pundartugg*. Narkotikarelaterade slanguttryck. Polishögskolan, 1997.

[SNPF96] Svenska narkotikapolisföreningen: *Basfakta om narkotika*, Göteborg, Sweden 1996.