Published in LSP & Professional Communication Volume 2, Number 1, April 2002 - ISSN 1601-1929
© DSFF / LSP Centre

REPORTS:

Multilingual Police Communication

LinguaNet now installed in 9 countries

Report on an EU Language Engineering project and its present application in police communication across European frontiers and languages

Inge Gorm Hansen & Henrik Selsøe Sørensen
Edward Johnson (Visiting professor at CBS, 1998-2000)
Copenhagen Business School, Denmark

1. Project background and recent developments

Since 1995, the language of operational police communication has been the focus of research for a team of LSP researchers from the Faculty of Modern Languages at Copenhagen Business School. The team has participated in two major EU projects: Test-Bed LinguaNet and Sensus, both dealing with language engineering related to law enforcement, the overall aim being to combat cross-border crime. This report will focus mainly on the language components of LinguaNet.

Basic project facts are shown in the fact box. Thus, CBS provided linguistic elements of the project; Prolingua provided a multilingual prototype and system software; the University of Bordeaux supported the linguistic effort; the University of Leuven set out the laws governing cross-border communication; Philips experimented with digital radio communications; the Judge Institute in Cambridge supported the awareness of the project and a multinational team of police officers in six countries was co-ordinated by the Kent County Constabulary.

More information is available at the project web sites, cf. below.

Facts about Test-Bed LinguaNet
EU programme: Telematics Applications
Programme domain: Language engineering
Budget, total: 2.4 million ECU
EU contribution: 1.4 million ECU

Participants
Cambridge University, United Kingdom
Copenhagen Business School
Kent County Constabulary, United Kingdom
Leuven University, Belgium
Philips CE BV, Netherlands
Prolingua Ltd, United Kingdom
Université Bordeaux II, France.

Main Project coordinator
Edward Johnson, Prolingua Ltd, Cambridge UK
CBS project coordinators
Henrik Selsøe Sørensen
Inge Gorm Hansen
CBS Research assistants
Bianca Hede, Hanne Erdman Thomsen,
Annemette Ruding, Ann June Sielemann

Active front-line police units communicate not only across national borders and language barriers but also across different national legal and penal systems involving complex domains of knowledge and highly specialised subject areas. These include drug trafficking, money laundering, organised crime, fraud, violent crime and economic crime, rules of criminal procedure to mention but a few. Police officers in one country and their counterparts abroad may have a common understanding of a wide range of professional issues, but do not necessarily possess sufficient knowledge about law, procedures and practices in the countries they interact with. Police communication is "knowledge intensive" and there is a need for transfer of highly specialised knowledge across languages and cultures.

LinguaNet provides police units with real time, language assisted, electronic, cross-border communication. Built in response to the relaxation of European frontier controls, LinguaNet offers a secure internet-independent link for routine communication between operational police units in different countries.

The system transfers and translates messages containing formatted controlled text segments dealing with identification and checking of persons, vehicles, bank cards, firearms etc., so that they can be read in any of the interface languages. In addition, the system allows machine translation of incoming free text segments. Images e.g. from digital cameras and sound files may also be integrated directly into the messages. Most carrier technologies can be used to transmit the messages from fixed or mobile terminals: ISDN, PSTN, X400, GSM, ADSL.

Used in concert with national bureaux of Interpol, LinguaNet is an additional tool to help police combat cross-border crime.

Originally, the project grew from an academic orientation, which coincided with an operational requirement. Operationally, no fast, safe and multilingual provision for cross-border police communication existed in a Europe busily engaged in dismantling its internal frontiers. Academically, no machine translation system could deliver translation of a quality sufficient for mission critical communication where an error might result in the loss of a life.

LinguaNet first began in 1992 as an English and French police and emergency prototype for services working at the Channel Tunnel. Beginning with bilingual cross-border emergency messages for the Channel Tunnel between France and England, LinguaNet was later configured for a range of routine police messages between a small number of countries in the region. It has since evolved and expanded.

In 1995, the European Commission DGXIII Framework IV Programme supported a project called Test-Bed LinguaNet; a three year effort to solve issues related to this and other developments. The project also allowed the further development of the operational system. Since 1999, the system has been developed further and is now commercially available. It has been tested, in front line conditions, between nine countries over a four year period. It performs reliably and consistently and a growing network of connections is established. We should distinguish here between the operational system which is implemented and the whole group of methodologies and concepts explored during the Test-Bed LinguaNet project. In several cases academic lines of enquiry went much further than was necessary for immediate system development. These lines of enquiry are discussed therefore, where appropriate, independently of the implementation.

The messaging system is used between nine countries and presently runs in eight languages. There are approximately fifty operational sites, most of which are port and border locations as well as central information bureaux. Forty-five of these are operational police sites which use the technology for daily communication, and the officers concerned meet up twice yearly at user group meetings.

At the same time, resources which may be deployed in further enhancements and academic research have been produced. These include the CBS collection of multilingual police terminology and the methodology created to assemble it, the results of experimentation with proprietary machine translation by CBS, studies of text/graphics integration and broad band radio transmission.

The LinguaNet Project has thus provided the opportunity not only to engage in computational, linguistic and communications research but to do it in a real world context and for a real world outcome. In addition to published papers, the Test-Bed LinguaNet team reports progress in highly unconventional terms:

- Stolen Vehicles worth over 5 million Euro recovered from containers at Felixstowe
- Drugs gang intercepted on the Spain/France frontier traced to Manchester
- Abducted child recovered in Holland
- Relative peace at the World Cup football matches at Lens France
- Stun gun attacker in Berlin found in Birmingham.

2. User involvement and user requirements

One of the most challenging aspects of the project is that research has been carried out in close co-operation with a user group - a multinational team of police end users in six countries coordinated by the Kent County Constabulary. The ongoing dialogue with the users has made the project “interactive”. The need for researchers to respond to user requirements and test results against the demands of the user community has demonstrated the value of user involvement and challenged existing working methods and methodologies.

A formal user requirement study was initiated in the early stages of the project. Contrary to many of the researchers' expectations, extremely practical considerations such as low price, low training implications, simplicity of operation, were highest on the list of priorities and not advanced technical solutions. Salient features of the user requirements which guided the system build are set out in the following list:

General / technical requirements

safe	- internet independent, encryption option
reliable	- low maintenance implications
point to point	- initial system to be server-independent
easy to install	- standard software platform, standard hardware
easy to use	- able to be used by non-technical staff
portable	- able to run on portable PCs via radio links
low training costs	- cheap to purchase and to run

transferable to multi-agency multi-national disaster senarios
able to use available connectivity
able to be upgraded (functionality)
able to be expanded (languages and sites)
able to carry graphics and sound files

Language requirements

interfaces in all user languages
user specified ***templates*** for operational messages
translation modules for controlled text and free text segments
creation of multilingual ***police lexicons and databases***

Some of the user requirement contained non technical issues which had a direct bearing on the technical details of the system build. A good example of this was the need for a thorough research study of existing legal provisions for data security and data interchange between the police forces of Europe. This included a legal study by the University of Leuven[1]. Another interesting feature was the requirement that direct connectivity to national criminal databases be specifically excluded for security, legal and data protection reasons.

To sum up, the system had to provide fast safe and direct communication between police units of any number of countries. It needed interfaces in multiple languages and the capability to translate messages about the basic topics used in real time cross-border police operations between those languages with full confidence. It would permit users to utilise less-than-perfect proprietary machine translation programs for gist translations of incoming free text. Further, dictionary look-up facilities and retrieval of police relevant information would be provided. The system would be capable of operating on or off the Internet, independently of service providers, independently of otherwise incompatible national police computer systems and permit direct exchanges between both static and mobile terminals.

[1] Van Outrive et. al. (1997): Legal Analysis (Final Report), University of Leuven Belgium.

3. Meeting user requirements

The remainder of this report will focus on the efforts to fulfil certain aspects of the language requirements.

A unique resource available to the project team was the collection of a corpus of operational language comprising i.a. messages exchanged between regional police units and between different branches of Interpol. Thus corpus analysis and thorough studies of procedures and linguistic features in police communication constituted the base of the selection of data elements in all languages. In response to the requirements set out by the users and as a result of the analyses, the three-layer model described below was developed as well as system interfaces in all user languages.

3.1 User specified templates and language interfaces

The multilingual corpus of police messages was analysed and relevant high frequency data elements identified and assembled in templates with pick-lists that constitute the backbone of the system. Identification and selection of data elements to populate the system was made by linguists in collaboration with police officers, and the resulting data elements were integrated into pick-lists as described below. Data elements, terminology, and factual police information were stored and managed in user-friendly databases cf. 3.3.

The **corpus analysis** and user surveys conducted during the user requirement stages of the project elicited the most commonly occurring message types. Examples of these are:

- wanted or missing person checks and replies
- vehicle checks and replies
- accident situation reports and updates
- bankers' cards enquiries and replies
- composite messages uniting persons, vehicles, drugs, firearms etc.
- requests and replies to telephone subscriber checks
- requests for and replies to address checks
- enquiries relating to small boats and movements thereof
- enquiries relating to light aircraft and movements thereof
- enquiries relating to companies.

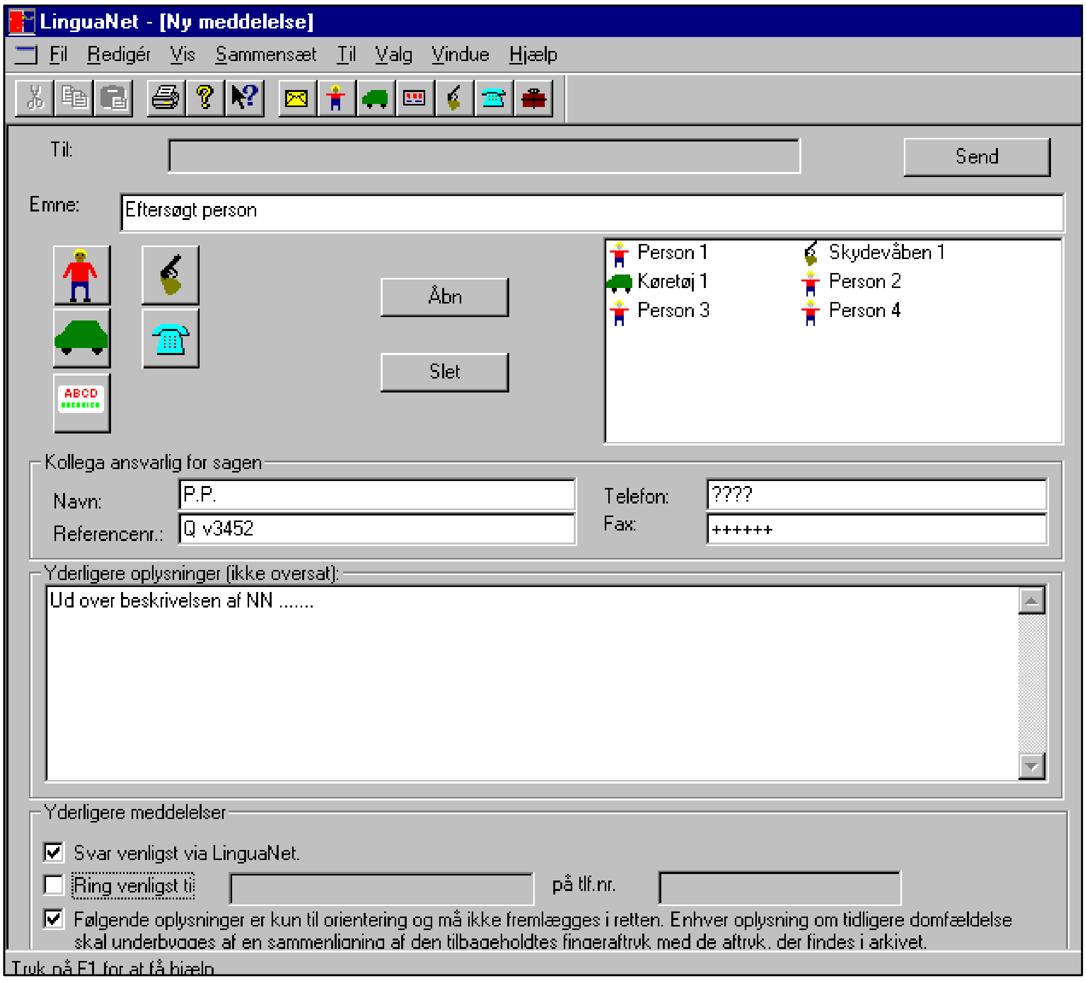
Data elements drawn from the corpus were analysed, compared to existing multilingual forms and guidelines (e.g. Interpol) and eventually assembled in multilingual templates representing a standard suited to LinguaNet. Thus, a number of templates were created for message types frequently occurring in cross-border police communication. Person description is a case in point. The below list shows part of the person description template covering eyebrows in three languages. The notation reflects the postion of each element in the knowledge model generated on the basis of the information extracted from the corpora.

LinguaNet standard description of eyebrows			
1.5.4-1.1-1-2	**øjenbryn**	**eyebrows**	**sourcils**
1.5.4-1.1-1-2.1	*(form)*	*(shape)*	*(forme)*
1.5.4-1.1-1-2.1.1	buede	arched	arqués
1.5.4-1.1-1-2.1.2	retlinede	straight	rectilignes
1.5.4-1.1-1-2.1.3	sammenvoksede	joining	réunis
1.5.4-1.1-1-2.2	*(dimension)*	*(dimension)*	*(dimension)*
1.5.4-1.1-1-2.2.1	korte	short	courts
1.5.4-1.1-1-2.2.2	lange	long	longs
1.5.4-1.1-1-2.2.3	brede	wide	larges
1.5.4-1.1-1-2.2.4	smalle	narrow	étroits
1.5.4-1.1-1-2.3	*(fylde)*	*(volume)*	*(volume)*
1.5.4-1.1-1-2.3.1	tynde	thin	minces
1.5.4-1.1-1-2.3.2	tætte	dense	épais
1.5.4-1.1-1-2.3.3	buskede	bushy	en broussaille

Each template consists of a set of predefined fields with picklists. Picklist values represent a selection of terms within the domains covered in the Test-Bed LinguaNet project. The automatic conversion between languages has two important features:

1) users at any site have a choice of interface languages and may switch from one to another. The interface language is usually determined by the language spoken in the country where the system is set up, e.g. English interface language in the UK, Danish interface language in Denmark etc.

2) the picklists match the interface language. This means in practice that a message sent from the UK in English is received in Danish in Denmark, in French in France etc.


LinguaNet - [Ny meddelelse]
Fil Redigér Vis Sammensæt Til Valg Vindue Hjælp
Send
Åbn
Slet
ABCD
Person 1
Skydevåben 1
Køretøj 1
Person 2
Person 3
Person 4

Til:

Emne: Eftersøgt person

Kollega ansvarlig for sagen

Navn: P.P.

Referencenr.: Q v3452

Telefon: ????

Fax: ++++++

Yderligere oplysninger (ikke oversat):

Ud over beskrivelsen af NN ....

Yderligere meddelelser

☑ Svar venligst via LinguaNet.

□ Ring venligst til på tlf.nr.

☑ Følgende oplysninger er kun til orientering og må ikke fremlægges i retten. Enhver oplysning om tidligere domfældelse skal underbygges af en sammenligning af den tilbageholdtes fingeraftryk med de aftryk, der findes i arkivet.

Tryk på F1 for at få hjælp

This screen dump shows the first page of a set of templates for a composite message covering an enquiry about persons, vehicles, stolen credit cards etc. In this case the interface and the user language is Danish. The icons signal that message refers to four armed persons in a car. To describe one of the persons, a click on the appropriate icon displays a page with templates for person description. Here, the sender may click the appropriate fields in order to describe the wanted person.

An important feature of this methodology is that it is fairly easy to add new languages. The analysis will have to be made for the new language only and results matched against existing data elements. Minor adjustments may have to be made in order to align and standardise data elements when a large number of languages are involved. However, adjustments will always be checked against and related to a controlled knowledge structure.

3.2 Translation modules for controlled text and free text segments

The preformatted messaging component can only process the most common routine messages. Although new messages may be analysed and added on a permanent basis, it is impossible to fulfil all user requirements related to a specific inquiry. Future candidates for standardised templates could be description of the crime scene and Modus Operandi. However, law enforcement officers often need to exchange case-specific information which cannot be standardised, and for that purpose free text fields are available in the system. This created the need for embedding a machine translation system allowing the text written in the free text field to be automatically translated.

It was therefore proposed to make one or more proprietary machine translation systems available to LinguaNet users. After consulting the user group, the following requirements were specified for the machine translation component:

- reliability in an operational context
- robustness
- must not presuppose linguistic expertise at the user level
- possibility to add customised dictionaries
- PC / laptop-based (given that LinguaNet uses point-to-point communication for security reasons).

At CBS, a test involving a user evaluation of the French => English module of SYSTRAN© was carried out with a view to establishing its potential in relation to the LinguaNet user group. SYSTRAN was applied in a small three-step experiment:

1. Segments of unedited corpus text were submitted to the standard SYSTRAN French-English module without correction of spelling mistakes etc.
2. A spell and grammar checked version of the same segments was submitted to SYSTRAN standard module
3. Finally, the checked texts were translated using a selection of SYSTRAN domain specific glossaries and relevant text type options plus a customised project dictionary.

On the input side, the availability of an interactive spell- and grammar-checking facility turned out to be crucial to ensure the quality of the source text to be translated. Some examples below illustrate some of the translation problems caused by typos.

A test panel was asked to read the translations from French into English and decide whether they

a) understood the messages from just reading the translated texts
b) found that they would act on the basis of the translation alone under mission-critical circumstances.

The user test concluded that only step 3 was a viable solution. According to the users, step 3 translations were comprehensible, but police officers would not rely on the translations alone under mission-critical circumstances. However, they found that the translations were by and large good enough for deciding whether or not to (urgently) call in a human translator.

The LinguaNet approach to using embedded MT was on the whole found useful provided that

- MT be available solely for translation of free text elements of an incoming message only as a supplement to controlled content transfer.
- the sender runs a spell and grammar check before transmitting a free text message; this could obviously not be done by the recipient who does not know the source language.
- it is the user in receipt of a free text message in a language he/she is not familiar with who submits the text to MT.
- any raw machine translation output be clearly flagged as 'raw MT output' thereafter.

A SPECIAL PROBLEM: TYPOS and MACHINE TRANSLATION

Spelling mistakes and typos may cause fatal translation problems in MT. Some are not detectable through spell-check: 'no body' (instead of 'nobody') was translated into French by 'aucun corps' (no corpse).

Others are detected but may still cause problems. Thus 'colleague' was spelled in the following ways in a corpus of 5000 messages, the number referring to frequency:

SINGULAR

57 colleague

2 colleages

6 colleaque

1 colleauge

3 colleauqe

22 collega

3 collegae

6 college

17 collegue

PLURAL

208 colleagues

7 collegas

1 colleuages

4 collegeas

3 colleaques

1 colleauges

The variation is quite impressive. The six occurrences of 'college' (should in all cases have been 'colleague') are not picked up by a spell-checker. If automatic spell-checking was used, mistakes like 'collegeas' would be changed to 'colleges' not 'colleagues'. English speakers who read 'college' instead of 'colleague' may easily understand the sentence anyway, but French policemen who read a translation which has 'université' instead of 'collègue' are not likely to detect the original error. These examples prove demonstrate the complexity of the problem of using machine translation under operational circumstances.

3.3 Multilingual police lexicons and databases

In addition to the messaging components and the templates, the Test-Bed LinguaNet results comprise a large multi-lingual knowledge and terminology repository. The compilation and structuring of the repository was headed by CBS in collaboration with police units across Europe. The main objective of the repository was to give police officers easy access to cross-cultural domain specific information as a supplement to the relatively restricted standardised knowledge integrated in the messaging system. At the same time the repository was the backbone of knowledge and terminology management in the project. The CBS repository comprises

- a person description database
- a Justice and Home Affairs database
- a drugs database
- a police facts database
- a casualty registration database

with a total of almost 10,000 concepts and around 35,000 terms including graphics.

Person description database

Person description is one of the most vital elements in police work. In recognition of this and in anticipation of LinguaNet extensions to many more languages a comprehensive set of data elements which can be retrieved for the purpose of form filling, report writing, translation, has been created. Data elements are organised in systems of concepts.

On the basis of source material provided a hierarchical system of concepts has been elaborated, the focal point being the description of the human head and face. One of the advantages of drawing up a system of concepts is that ‘gaps’ in the overall coverage are disclosed, and the hierarchical set-up gives an overview of the structure of the particular subject. Terms have been organised in a systematic list including position numbers derived from the system of concepts.

The database resulting from the data element engineering contains

1. Comprehensive person description terminology in three languages (Danish, English, French),
2. Systems of concepts in the three languages,
3. Pictures of physical features, and links between records that enable users to track knowledge about a certain term and go from broad to narrow concepts. If, for example, the user wants to describe a head more closely, he/she may look up the term head and gain access to sets of narrower terms.

The record for head has links to records for shape of the head en face and shape of the head in profile. The record has the relevant terminology for describing a head as well as graphics and a link to the appropriate system of concepts.

The Justice and Home Affairs Database

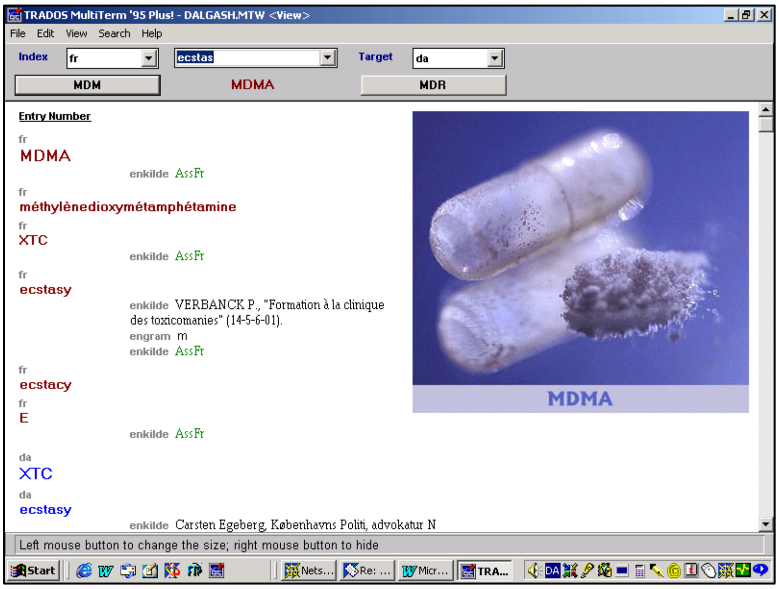
The Justice and Home Affairs Database consists of terminology retrieved from the European Commission's Customs Information System database. The multilingual information included comprises subject areas of the EU treaty section VI, article K, i.e. asylum policy, immigration policy, police co-operation (combating drug trafficking, terrorism and organised crime).

The Justice and Home Affairs terminology covers 9 EU languages: Danish, Dutch, English, French, German, Italian, Portuguese, Spanish, and Swedish. Cross-references in the database allow the user to easily access associated concept entries.

Drugs Database

Internet resources concerning police matters have proved to be a useful knowledge source in the project, e.g. Internet sites concerning wanted people, stolen art, missing children, drugs, etc. CBS created a multilingual knowledge base on drugs with their street names, pictures, etc.

The information contained in the database comes mainly from official sources in English, French, German, Spanish, and Danish.


TRADOS MultiTerm '95 Plus! - DALGASH.MTW <View>
File Edit View Search Help
Index fr
ecstas
Target da
MDM
MDMA
MDR
MDMA
Start
Nets...
Re: ...
Micr...
TRA...
DA

Entry Number

fr
MDMA
enkilde AssFr

fr
méthylènedioxymétamphétamine

fr
XTC
enkilde AssFr

fr
ecstasy
enkilde VERBANCK P., "Formation à la clinique des toxicomanies" (14-5-6-01).
engram m
enkilde AssFr

fr
ecstacy

fr
E
enkilde AssFr

da
XTC

da
ecstasy
enkilde Carsten Egeberg, Københavns Politi, advokatur N

Left mouse button to change the size; right mouse button to hide

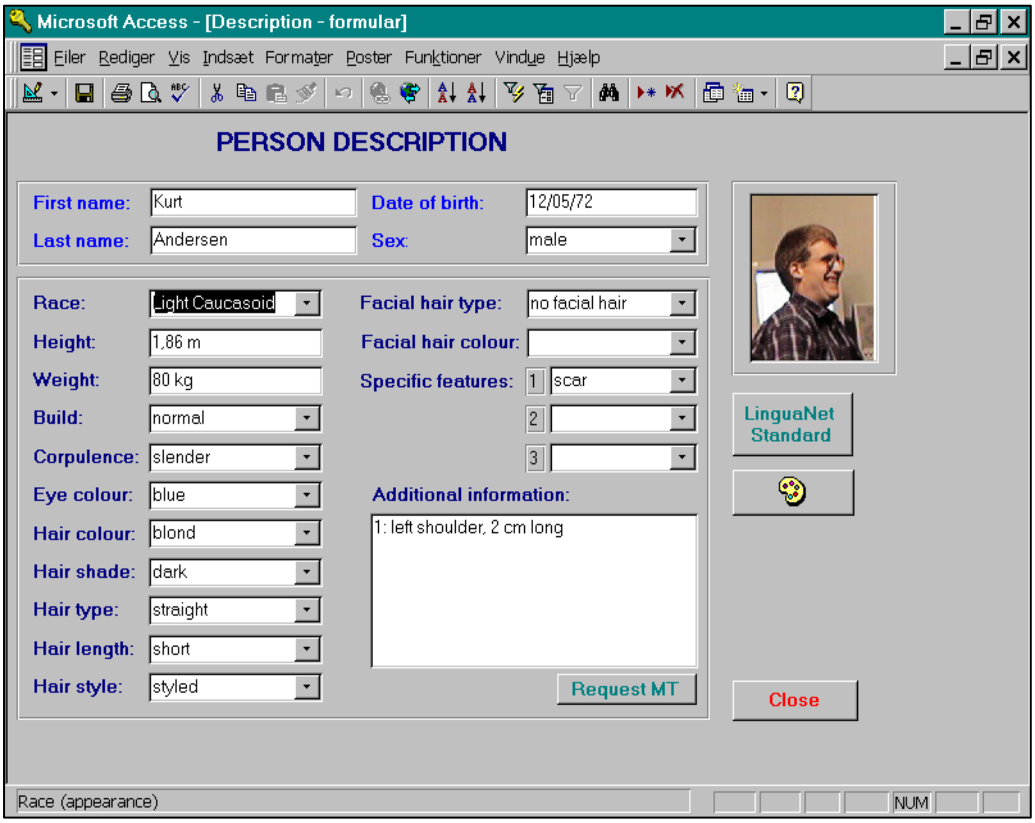
Police Facts Database

This corpus also forms part of the Test-Bed LinguaNet Knowledge Base. The corpus comprises terminology covering policing and rescue terms originating from a number of sources, the main one being a publication in Danish, English, French and German edited by the National Commissioner in Denmark. It gives an introduction to functions, duties, organisation, personnel, etc. of the Danish Police Service containing extensive terminology for description of police including text, concepts, graphics and pictures. This material has a clearly defined source language, which is Danish. This is reflected in The Test-Bed LinguaNet Knowledge Base in that the English, French and German terminology is used to describe the Danish police services. The aim of this work was to develop a methodology applicable to other languages.

Multilingual Casualty Registration

CBS has created a flexible modular system for registering and retrieving incident and person data in connection with a major incident. It does not, so far, form part of the LinguaNet Windows version, but is a prototype for a prospective casualty format.

One of the forms - the Person Description Form - is the result of extensive terminological work on assembling data elements and building concept systems for person description. The information categories represented have been selected by comparing and contrasting existing forms and attempting to satisfy both the need


Microsoft Access - [Description - formular]
Filer Rediger Vis Indsæt Formater Poster Funktioner Vindue Hjælp
LinguaNet Standard
Request MT
Close
NUM

PERSON DESCRIPTION

First name:	Kurt	Date of birth:	12/05/72
Last name:	Andersen	Sex:	male

Race:	Light Caucasoid	Facial hair type:	no facial hair
Height:	1,86 m	Facial hair colour:	
Weight:	80 kg	Specific features:	1 scar
Build:	normal		2
Corpulence:	slender		3
Eye colour:	blue	Additional information:	
Hair colour:	blond		1: left shoulder, 2 cm long
Hair shade:	dark		
Hair type:	straight		
Hair length:	short		
Hair style:	styled		

Race (appearance)

for quick identification, where only a few information categories are filled in, and a more thorough person description with more details. Picklists reflect the work done on person description terminology.

3.4 Interfaces in all user languages

At present, LinguaNet has user interfaces in Danish, Dutch, English, French, German, Italian, Polish, Portuguese, Spanish and Swedish.

4 Future

In this paper we have focused on the language problems in cross-national police communication and the language resources and data elements developed for LinguaNet. The LinguaNet user community has grown significantly since the conclusion of the EU project. The users find the system as a whole user-friendly and robust and consider the following features developed in the LinguaNet project particularly useful:

a) Easy access to stored and pre-translated data elements supplemented with free text elements.
b) Information retrieval in one or several languages from knowledge bases, lexicons, terminological resources.

c) Interaction between text, pictures and speech.

According to the users, another important advantage has been the creation of trust and confidence between colleagues who speak different languages but need to work closely together.

Future developments could comprise enhancement of the present system with such technologies as speech to text and vice versa, deeper integration of graphics, seamless access to professional databases.

Cross-border communication is in its very nature multilingual and with the disappearance of national borders in Europe there is an increasing need to overcome existing language barriers not only in police work but also in business communications, medical communications etc. The LinguaNet system structure and methodology can be applied wherever the communicative requirements of the participants can be engineered as a series of suitably controlled transactional events using controlled language. The system is easy to expand so as to cover new languages and new domains. The most recent LinguaNet user is the British Immigration Authority, and this may signal the starting point of a new important development for the LinguaNet system.

Futher information
Edward Johnson, Prolingua Ltd, United Kingdom
E-mail: prolingua@prolingua@co.uk

Inge Gorm Hansen
E-mail: igh.eng@cbs.dk

Henrik Selsøe Sørensen
Email: hss.fra@cbs.dk

LINGUANET WEB SITES

LinguaNet featured as a showcase project	http://www.prosoma.lu
LinguaNet within a EuropeanCommission language engineering project	http://www.hltcentral.org/ (search for LinguaNet)
LinguaNet main website	http://hermesdoc.lib.cbs.dk:80/departments/fir/linguanet/
LinguaNet brochure (1996)	http://www.Prolingua.co.uk/brochure/contents.html

***

ABSTRACT

Multilingual Police Communication

Report on an EU Language Engineering project and its present application in police communication across European frontiers and languages

Inge Gorm Hansen & Henrik Selsøe Sørensen
Edward Johnson (Visiting professor at CBS, 1998-2000)
Copenhagen Business School
Denmark

The LinguaNet police communication system is significant for four reasons. Firstly, the number of frontier locations using LinguaNet has increased greatly from the original pair of just two police offices in Kent and the Nord Pas-de-Calais. Secondly, the LinguaNet initiative became broader in scope than the Channel Tunnel projects and was eventually supported by the European Commission as a six nation Framework 4 Telematics project. Thirdly, LinguaNet is a provision which recognises the importance of first hand observations and the sharing of those observations between trusted law enforcement professionals across language barriers. Terrorism and organised crime cannot be defeated by techniques such as data mining and analysis of electronic footprints alone. Fourthly, LinguaNet, as a project, has metamorphosed from a custom-built police communications mechanism to a living experiment in cross border co-operation which is running still. Much is being learned from the experiment which, when properly recorded and analysed, can provide vital data for future developments. This paper focuses on the contributions by CBS to the LinguaNet project; these contributions were concentrated on linguistic analysis, terminology extraction and terminology management as well as integration of embedded machine translation.

***