Journal of Hazardous Materials, 27 (1991) 213–230
Elsevier Science Publishers B.V., Amsterdam

Project Report

Machine translation for chemical safety information

J. Takala, J. Pesonen, L. Kulikov[1] and H. Jäppinen[1]

Asian Regional Chemical Safety Information Project,
International Occupational Safety and Health Information Centre, CIS, International Labour Office, CH-1211, Geneva 22 (Switzerland)

(Received May 17, 1990; accepted in revised form December 7, 1990)

Abstract

Hundreds of thousands of Chemical Safety Information Sheets (Material Safety Data Sheets) exist in main world languages. A few data bases exist in a machine readable form. In order to benefit from the bulk of this information a special microcomputer based translation system was developed, which translates the information sheets from the Finnish mainframe data base KETURI, Finnish Register of Chemical Safety Data Sheets, containing ca. 45,000 data sheets, into English. The translated sheets will be further used by the seven participating developing countries once they have prepared their own chemical safety data sheets in their respective languages. The intermediate results, the English language chemical safety data sheets may be processed and disseminated in various forms: as word processor files, microcomputer data bases, on microfiche, and on CD-ROM (Compact Disc - Read Only Memory). The initial results, after the processing of ca. 1000 sheets, are encouraging. Although the language is relatively limited, the idea of using artificial intelligence in translating bulks of safety information is revolutionary and may have far reaching consequences. The project was established jointly by the International Occupational Safety and Health Information Centre, CIS, of the International Labour Office, ILO, and the Kielikone Project of the Finnish SITRA Foundation.

Introduction

In order to reduce the number of accidents and occupational diseases, as well as environmental spills, leaks, accidental releases, emergency situations and waste disposal problems caused by hazardous materials, there are naturally a number of different methods. The two key elements in the safe use of chemicals are adequate information on the inherent hazards and the proper use of these substances, and effective means of disseminating the information to workers and those persons responsible for their safety and health. Information on

[1]Finnish National Fund for Research and Development, (SITRA Foundation), Kielikone Project, P.O. Box 329, SF-00121, Helsinki, Finland.

0304-3894/91/$03.50

chemicals relating to health and the environment can be obtained from several sources: publications and scientific journals, bibliographic data bases, and factual data banks [1]. A distinct method, which has recently received a lot of attention, is that of using the Chemical Safety Data Sheets (or Material Safety Data Sheets as they are called in the North American Continent) as a vehicle to spread essential information on chemicals and chemical products. A Chemical Safety Data Sheet identifies a hazardous substance, its manufacturer or importer, its hazards to safety and health, and precautions to follow when using it [2]. However, CSDSs have also been criticized as being both too technical for many people and failing to address public concerns and questions [3].

In the USA there is a clear requirement, that the manufacturer must provide these sheets along with the chemical supplied [2]. Similarly, in Canada, the new WHMIS (Workplace Hazardous Materials Information System) sets the same standard [4].

An analogous system called TVATM (Identification and Labelling System for Substances Hazardous to Health) [5] was introduced in Finland in 1979.

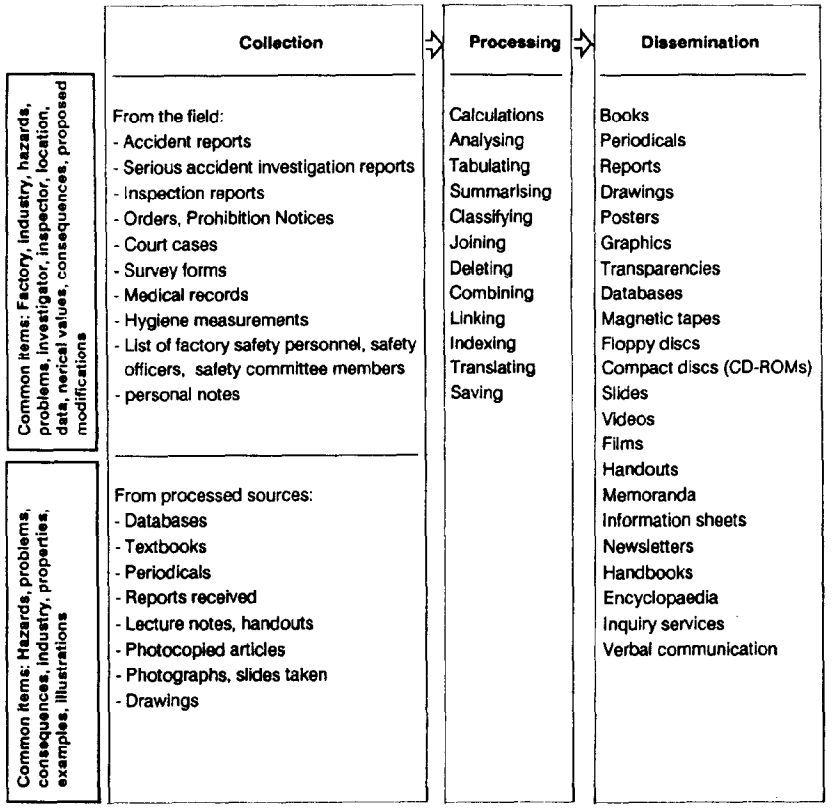

Fig. 1. Information activities in safety and health.

Since that date the Finnish manufacturers and importers of chemicals have been obliged to submit a full factual chemical safety data sheet to the suppliers and users of chemicals and another copy to the authorities, the National Board of Labour Protection. This body, after checking the quality of the information given on the data sheets, then established a mainframe factual data base presently containing ca 45,000 data sheets out of which more than 20,000 have been validated. In order to exploit this unique data base containing validated information on a huge amount of chemicals, a machine translation system from Finnish into English has been developed. The translated information is particularly aimed at meeting the immense need for practical safety information in developing countries. Large amounts of highly relevant data exist in industrialised countries, already collected, organised and processed, but in languages incomprehensible by the users in need in the third world. The dissemination of information is crucial, but in order to achieve its objective of reducing risks and hazards, the data must also be effectively communicated in a suitable way.

The theory of an information system for occupational safety and health and hazardous materials contains three distinct elements: the collection, processing and dissemination of this information as described in Fig. 1 [6].

Why chemical safety data sheets?

To complement the legislation of individual countries the International Labour Office has begun a task with a view to establishing an ILO convention, i.e. international agreement on "Safety in the Use of Chemicals at Work". One of the key elements of this present (May 1990) draft convention is the provision of information on Chemical Safety Data Sheets [7]. This means that these data sheets will be increasingly used when providing information on hazardous materials worldwide. There are certainly benefits in the use of this media as a vehicle for dissemination of chemical information (see Fig. 2).

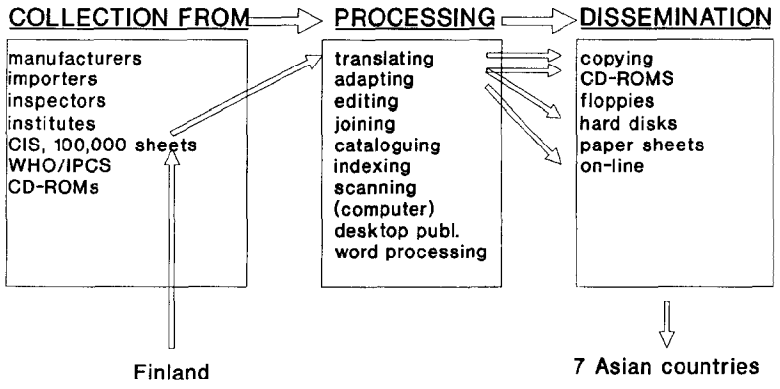

Fig. 2. Chemical information sheets: Machine translation and information process.

Existing solutions in chemical safety

Machine translation or computer aided translation as an idea has been discussed for some time and a few solutions have indeed already been developed. An example of large mainframe computer applications already developed is the SYSTRAN solution, which is designed for any type of document translation in the Commission of European Communities in Luxembourg [8]. This system is capable of accepting free text in several Community languages and producing a draft translation in a form normally accepted by word processing programmes. The text is then checked, improved and edited by human translators. Examples of machine translated safety and health texts, and test translations of CIS Abstracts are presented below in Fig. 3.

Although SYSTRAN is able to produce fairly high quality draft translations,

ORIGINAL

CIS 85-1954 Occupational exposure to carbon disulfide and vitamin B_6 deficiency (Exposição ocupacional ao sulfeto de carbono e hipovitaminose B_6). De Souza Nascimento E., Mídio A.F. *Revista brasileira de saúde ocupacional*, Apr.-June 1984, Vol.12, No.46, p.17-24. Illus. 64 ref. (In Portuguese)

A literature survey of the toxic action of carbon disulfide in exposed people, in particular of its inhibition of the pyridoxamine form of vitamin B_6, of its interference with tryptophan metabolism, and of the resulting excretion of xanthurenic acid. (45016)

MACHINE TRANSLATION

85-1954 CIS –(45016)
Exposition professionnelle à l'insuffisance de bisulfure et de vitamine B_6 de carbone
Portugais
Une enquête de littérature de l'action toxique du bisulfure de carbone chez les personnes exposées, notamment de son inhibition de la forme de pyridoxamine de la vitamine B_6, de son interférence avec le métabolisme de tryptophane, et de l'excrétion résultante de l'acide xanthurenic.

HUMAN TRANSLATION

CIS 85-1954 Exposition professionnelle au sulfure de carbone et hypovitaminose B_6 (Exposição ocupacional ao sulfeto de carbono e hipovitaminose B_6). De Souza Nascimento E., Mídio A.F. *Revista brasileira de saúde ocupacional*, avr.-juin 1984, vol.12, n°46, p.17-24. Illus. 64 réf. (En portugais)

Etude bibliographique concernant l'action toxique d'une exposition au sulfure de carbone qui se manifeste surtout par son pouvoir inhibiteur sur la forme pyridoxamine de la vitamine B_6, par son interférence avec le métabolisme du tryptophane et par l'excrétion corrélative d'acide xanthurénique. (45016)

Fig. 3. Example of a chemical information sheet.

it has certain weaknesses: (1) The system is only accessible through a limited number of mainframe computers; (2) It cannot be modified to satisfy translation needs of special nature i.e. in order to use a full range of specialised occupational safety and health terminology: (3) It requires professional computer programmers and cannot be used directly by for example, safety professionals.

Another large mainframe computer application, and in this case designed particularly for the preparation of new chemical safety information sheets in various languages, is the system developed by DuPont de Nemours International S.A. It is not directly a machine translation system, but an instrument using elements of machine translation, when new sheets are drafted. It is based on fixed phrases, which may be combined in various ways. However, it has exactly the same deficiencies, if judged from the safety professional's point of view, as the SYSTRAN.

A third solution is the translation of the Registry of Toxic Effects of Chemical Substances (RTECS) [9], of the US National Institute of Occupational Health, NIOSH. The RTECS database was originally created with a set of standard codes in order to reduce the amount of keying in work and mass storage size. The coded system is available on-line and on compact disc, the CHEMBANK, CD-ROM [10]. In order to increase readability, the codes were translated back into full English phrases (of type: code - Engl - French) and later into an equivalent full French-text database by the Canadian Centre of Occupational Health and Safety (Fig. 4). These two full-text versions were then mounted on a Canadian CD-ROM, CCINFOdisc A1 [11].

A fourth solution, for machine translation in chemical and occupational

Fig. 4. RTECS - translation of data base.

Coded version	English version	French version
T/E n.y.rvd ihl-hmn LCLo: 1300 ppm/30M "Practical Toxicology of Plastics,"	INVESTIGATED EFFECT Toxic ROUTE OF ADMINISTRATION Inhalation ORGANISM OBSERVED Human TEST OR DOSAGE TYPE LCLO - Lowest published lethal concentration CHEMICAL AMT/CONC/DUR 1300 ppm/30M "Practical Toxicology of Plastics"	EFFET A L'ETUDE Toxique VOIE D'ADMINISTRATION Inhalation ORGANISME OBSERVE Humain GENRE DE TEST OU DE DOSAGE LCLO - Plus basse concentration létale publiée QUANTITE DE PRODUIT/ CONCENTRATION/DUREE 1300 ppm/30M "Practical Toxicology of Plastics"

safety and health information, is the TERMS file (15,000 records) of the International Occupational Safety and Health Information Centre, of the International Labour Office. In order to translate on a continuing basis main and secondary CIS descriptors, classifying and identifying related topics with a controlled vocabulary, a facet coding system was established [12–14]. These facet codes are arranged in a hierarchical mode grouping each descriptor into a larger community. Once these "international language" facet codes are linked to natural language descriptors and even longer expressions and phrases, the machine translation and validity checking is a simple task. In addition to the hierarchical facet codes all chemical and product names as well as synonyms are linked to the Chemical Abstract Service CAS numbering system, which provides for an unambiguous and internationally accepted name for practically any known chemical substance (more than 6 million registered). This system is used in a number of other solutions and especially in computer data bases where relations with different files may be established. Some 4000 CAS numbers have been used as the "chemical names" in the CIS Thesaurus and machine translation system. This system has been operational for several years in CIS and in many CIS National Centres all around the world and is currently being revised.

Example:

facet code	CAS No.	English descriptor	French descriptor
Cato	7664-93-9	SULFURIC ACID	ACIDE SULFURIQUE

Development of a new system for translation of chemical information

Background

The objective was to develop a machine aided human translation system in order to translate into English essential elements of the Finnish Register of Chemical Safety Information Sheets (Kemiallisten Tuotteiden Käyttöturvallisuusrekisteri, KETURI) with a microcomputer and to use this material for further translations into local languages in developing countries (using conventional methods). The complete data base is available, in Finnish, in magnetic form.

Machine translation is a difficult task if the text is not restricted any way [15]. However, there are certain cases in where a language is in fact restricted. Such sublanguages usually make the problem easier, but they can also add new difficulties.This occurs when the text is under limited subject domain and vocabulary (chemical safety) or where the text is written in telegraphic style: often short and incomplete sentences. Some words and phrases are also frequently repeated in the text.

Difficulties to be overcome

The difficulties of the natural language translations are, to mention a few, the huge number of different expressions having almost the same meaning, homographs, synonyms (especially those of chemical names), the different use of active and passive forms in English and Finnish, where the passive is often used as a polite and formal imperative, and other non-regular verb forms. The output must also be clearly understandable by non-specialist, non-native English speakers in the seven developing countries in Asia, where the results are expected to be put into practice in preparing local language chemical safety data sheets. These sheets are to be disseminated at the factory shop-floor level in a form such that users and those who are transporting or handling hazardous chemicals will understand the text without an extensive amount of training.

Problems related to quality of the original text are: misspelling in the original text, unconventional abbreviations, unorthodox grammatical forms, numerous individual expressions where hundreds of industrial hygienists have been involved in the preparation of the Finnish data sheets.

Limitations are also created by the technical environment. In order to keep the system controllable, affordable and easily transferable, the target was to create the full machine translation system with commonly available microcomputers, i.e. IBM AT compatible hardware and with the standard MS-DOS, disk operating system.

Description of the system

The system runs with IBM AT-compatible, or XT with 640 kb user memory, microcomputers with MS-DOS 3.0 or more recent versions. It needs at least 600 kbytes of main memory to run. The translation is divided into the following eleven parts:

(1) Separating the units to be translated.
(2) Morphological analysis and updating of the unilingual dictionaries
(3) Checking of general bilingual dictionaries and the special dictionary for chemical names and synonyms
(4) Checking of the directory for repeated free-text phrases with the pattern rulebase and fixed Safety- and Risk-phrases for warning symbols, labelling and identification requirements
(5) Translation using the grammatical rulebase
(6) Editing of the output of the English version by a human translator
(7) Creating the final English chemical data sheet forms with titles and field names.

The system has three dictionaries and two rulebases: a Finnish morphological dictionary, general and chemical Finnish–English dictionaries, rules for recognising commonly used phrases and rules for recognising grammatical structures.

(1) Separating the units to be translated

Firstly the system picks up from a Finnish data sheet file the data which will be transmitted to the English version. An entire data field is read and analysed. Some of it can be transmitted as it is (e.g. names, addresses, numerical values). The text is translated one unit at the time. A unit starts at the beginning of a field or where the previous unit ended. It ends at a full stop, comma, colon, semicolon, hyphen or bracket. The units translated vary from one word to entire sentences (Figs. 5 and 6).

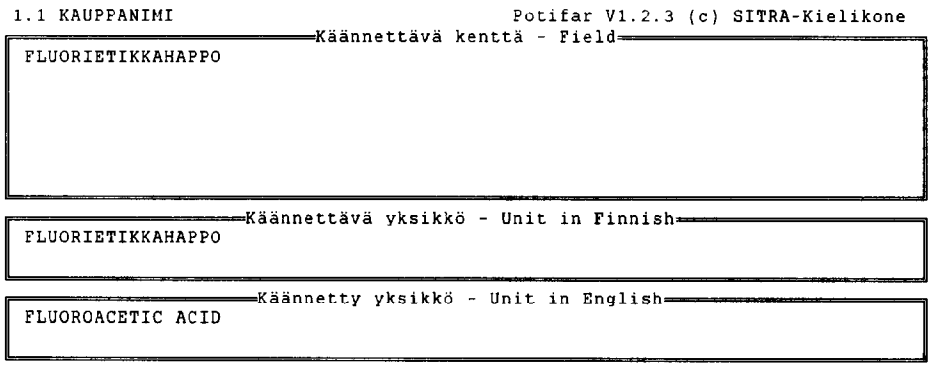

Fig. 5. Example of the KETURI data sheet file (blank) conversion.

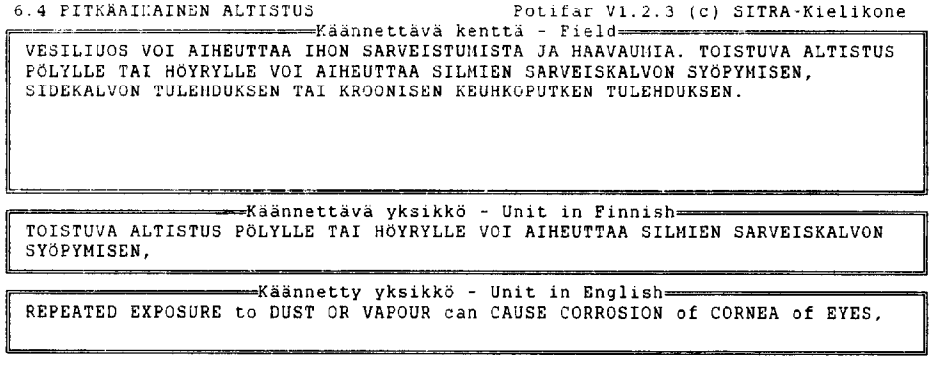

Fig. 6. Practical example of the translation.

(2) Morphological analysis

Each word form is analysed with MORFO, a morphotactic parser for Finnish words developed by the SITRA-foundation [16]. For each word form MORFO gives the grammatic interpretation and the basic lexicographical form. MORFO also detects misspelled or new words, and they can be corrected or added to the lexicon in a convenient manner. After the whole unit has been analysed it will be transmitted to the next phase.

The morphological analysis can be compared to the spelling checker of an advanced word processing programme; MORFO was in fact developed for such purposes.

(3) Checking the general dictionary

Before translating, the system checks that there is an English counterpart for each word in the unit. Missing words are added at this point, but changes can be made during translation. Special dictionaries have been established for nouns, chemical names and synonyms, verbs and verb forms, as well as other words.

(4) Phrase translation with the pattern rulebase

After morphological analysis every word in a sentence has an unambiguous grammatic interpretation. The next step applies hierarchical translation rules to the sentence. The rulebase of the system consists of two parts. The pattern rules recognise entire free-text phrases and give their English equivalents. The system first checks if any of these rules match with the unit. If they do, the phrase will be replaced automatically by its English equivalent (postediting is not needed in this case).

Similarly the Risk- and Safety-phrases used for standard labelling and identification of chemical containers and packages will be located within the text and directly translated. These R- and S-phrases were originally established by the Commission of the European Communities [17] and adopted, or being considered almost as such by a number of other countries (Nordic countries, Hong Kong, Thailand [18,19], etc.). These include more than 120 different standardised phrases, which are all coded and easy to detect and translate.

Examples:

R 23/24/25 = Toxic by inhalation, in contact with skin and if swallowed

S 36/37/39 = Wear suitable protective clothing, gloves and eye/face protection

When no pattern rules exist, general rules are applied.

(5) Translation using the general grammatical rulebase

General rules recognise grammatical constituents such as nouns, adjective phrases, noun phrases and verb phrases. The rules are context-sensitive, taking into account not only the surrounding words but also their grammatical forms. The building of constituents starts with Adjective Phrases (AP), goes through a series of Noun Phrase (NP) rules, then binds Verb Phrases (VP) and finally Preposition Phrases (PP). In addition to building constituents the rules translate their words (those which have not been translated so far). Word ordering is handled by numerical indexes, and a change in ordering can be effected by a new ordering of indexes.

When the morphological analysis using the unilingual Finnish lexicon is carried out, MORFO also notes down the forms of the words like case endings, plurals etc. These notes are used by the rulebases and converted into equivalent English forms, displayed in lower case characters in Fig. 6. The bilingual dictionary divided into groups according to the word class and used in the translation phase proper provides the plain English word, in upper case characters in Figs. 5 and 6. The human translator can modify both dictionaries, by adding new words to them or removing incorrect entries. The bilingual dictionary can also be used and edited independently from the translation task. The chemical vocabulary is separated from general words, and it forms a special lexicon.

Due to the microcomputer memory restrictions only the Finnish words of the bilingual dictionary are kept in the main memory. After a successful match the English equivalents are retrieved from the hard disk. This strategy reduces translation speed but allows the use of large dictionaries. The general vocabulary contains about 10,000 words in each category of words and the chemical lexicon will have another 10,000 entries.

(6) The working environment and tasks of the translator

When the translation work is running, the text to be translated is extracted from the data fields one field at the time. A whole field is displayed in a window to show the context to the human translator. Automatic translation is done one sentence (text unit) at a time and that text unit is displayed in another window. When the text unit has been translated the results are shown in a third window. The translator either accepts the translation, marks it for checking later and/or post-editing, or he can make corrections right away (Figs. 5 and 6). On average the speed of translating of one text unit is about two seconds. The corrected sentence can be saved as a standard phrase exactly as for individual words. After such saving the expression is always translated correctly without any need for confirmation. Thus the more sheets have been translated the less words and phrases need to be given or confirmed; the system works quicker, it "learns".

(7) Creating the final English chemical data sheet

After translation the information is in basic ASCII-format. A final English form with titles and field names is created by using a separate programme. The translator then checks the quality of translation and postedits the text with word processing software (WordPerfect 5.0). The labelling and identification symbols are also added with a WordPerfect 5.0 graphics facility after the symbols had been scanned with an image scanner. Further editing can be effected with Ventura desktop publishing software.

A model of the new translated CSDS is presented below (Fig. 7). Only basic editing has been done with a view to the text not having to be 100% correct in order to be understood by users. However, the outputs may easily be improved with any common text editor. In order to avoid confusion, those original data sheets, which did not have a CAS number(s), have been provided with these, thus improving the quality of the output.

Results

The machine translation system described above has been completed and some 1000 CSDSs have been translated. In addition to the "mass production" from the original file downloads, a special filter was created to translate individual data sheets whenever necessary. Thus a single or a few specially detected data sheets can be downloaded, e.g. using a modem connection from Switzerland to Finland and subsequently translated for a customer inquiring information about the substance. Some 30 particular data sheets on major hazard chemicals of the EC-list [21] have been translated for a data base on these substances. The target is to translate a few thousand of the most relevant data sheets.

However, the original target to translate the sheets in a batch mode was not reached – although it is technically easy, it would not be feasible quality-wise at this moment – due to following reasons:

(1) The quality of the original sheets is not homogeneous; some of them could easily be machine-translated in a batch mode, some others not; biggest problems are random spelling mistakes and unconventional and unpredictable abbreviations;

(2) A large amount of expressions that are grammatically incorrect;

(3) The number of phrases grows quickly as hundreds of individual industrial hygienists have probably been involved in the preparation of the data sheets – this has resulted in tens of different phrases to express one single and simple meaning;

(4) The MORFO programme does not recognise some of the specific terms; this would, however, improve as the system learns more;

(5) The disk operating system will at some point be the limiting factor as addresses of dictionaries will have to be kept in the user memory.

FS-20

NATIONAL BOARD OF LABOUR PROTECTION IN FINLAND
(machine translated data sheet, CIS-ILO, CH-1211 Geneva 22)
*** no liability accepted by CIS-ILO *** 1989 ***

NO: 05154-0 DATE: 02.01.1981

1.1 TRADE NAME: TETRAHYDROFURAN

1.2 USE: LABORATORY CHEMICAL

1.3 MANUF/IMPORTER: SAI-LAB C/O SUOMEN TUKKUKAUPP. LIITTO
ADDRESS: FABIANINKATU 23
00130 HELSINKI 13 TEL:90-11806

2.1 TOXICITYCLASS: 2 2.2 FLAMMCLASS: 1 2.3 TRANSPORTCLASS: 3
2.4 UN-NUMBER: 01165 2.5 CARCINOGENS: ASA- ASA- ASA- ASA-

2.6 WARNING LABELS: F HIGHLY FLAMMABLE
XI IRRITANT
R11 HIGHLY FLAMMABLE.
R19 MAY FORM EXPLOSIVE PEROXIDES.
R36 IRRITATING TO EYES.
R37 IRRITATING TO RESPIRATORY SYSTEM.
S16 KEEP AWAY FROM SOURCES OF IGNITION - NO SMOKING.
S29 DO NOT EMPTY INTO DRAINS.
S33 TAKE PRECAUTIONARY MEASURES AGAINST STATIC DICHARGES.

---

3 SUBSTANCES HAZARDOUS TO HEALTH:

1 TETRAHYDROFURAN
CAS:109-99-9*%100.
LDL0*03000 MG/KG (ORAL, RAT) MAC = 200 CM3/M3
= 590 MG/M3 FLAMMABLE, VOLATILE, LIQUID HARMFUL TO HEALTH.

4.1 BOILING POINT: BP:*00064CE

4.2 MELTING POINT: MP:*-0108CE

4.3 VAPOUR PRESSURE: VP:*20,0KPA (20 CE)

4.4 SOLUBILITY IN WATER: FULLY SOLUBLE

Fig. 7. Machine translated data sheet of tetrahydrofuran.

4.5 DENSITY:	890 G/DM3
4.6 EVAPORATION RATE:	ER:*5,0
4.7 pH-VALUE:	PH:
4.8 PHYSICAL STATE:	COLOURLESS LIQUID. ETHER RESEMBLING ODOUR.
5.1 FLASHPOINT:	FP:*-0017CE
5.2 FLAMMABLE LIMITS:	LIL:*001.5 HIL:*012%
5.3 AUTOIGNITION TEMP:	AIT:*0260CE
5.4 REACTIVITY:	WHEN HEATED DECOMPOSES FORMING TOXIC GASES. CAN FORM EXPLOSIVE PEROXIDES. CAN REACT VIOLENTLY WITH OXIDANTS.
6.1 ROUTES OF EXPOSURE:	EVAPORATING SOLVENT. SOLVENT VAPOURS ARE DANGEROUS.
6.2 LOCAL EFFECTS:	SOLUTION IRRITATES SKIN. SOLVENT VAPOURS IRRITATE EYES AND MUCOUS MEMBRANES.
6.3 SHORT-TERM EXPOSURE:	SOLVENT VAPOURS CAUSE SMARTING IN EYES AND RESPIRATORY SYSTEM, HEADACHE AND VERTIGO.
6.4 LONG-TERM EXPOSURE:	REPEATED SKIN CONTACT OF SOLUTION CAN CAUSE ECZEMA.
7.1 SAFETY PRECAUTIONS:	SEPARATE FROM SOURCES OF IGNITION - SMOKING PROHIBITED. PREVENT FORMING OF STATIC ELECTRICITY. RECOMMEND PROTECTIVE GLOVES (ACCIDENT SITUATIONS). AVOID DIRECT SUNLIGHT.
7.2 FIRST AID:	TAKE INTOXICATED PERSON TO FRESH AIR. RINSE EYES AND SKIN WITH WATER.
8.1 STORAGE:	STORE PROTECTED FROM LIGHT IN A STORAGE FOR FLAMMABLE LIQUIDS.
8.2 CORROSIVENESS:	DISSOLVES RUBBER.
8.3 SPILL CONTROL:	ABSORB WITH PAPER OR SOMETHING ELSE , EVAPORATE IN FUME CUPBOARD AND BURN IN SAFE PLACE.
8.4 ENVIRONMENT HAZARDS:	DO NOT RELEASE TO SEWAGE SYSTEM.
8.5 WASTE DISPOSAL:	BURN UNDER CONTROLLED CIRCUMSTANCES.
8.6 FIRE PRECAUTIONS:	EXTINGUISH FIRE WITH WATER OR IN ANY OTHER WAY. COOL CONTAINERS EXPOSED TO FIRE.

Fig. 7. Continued.

Already at this moment the system has proved economical; the price for a single sheet translation is about 30 ECU (US$40) while the expenditure for human translation would have been more. Furthermore, the outputs are readily available in both paper and magnetic optical form for inclusion into a new data base. The expenses per translated sheet will be significantly further reduced once the amount of translated sheets increases.

As side products, a number of bilingual dictionaries have been created, as well as a most valuable collection of standard phrases related to chemical safety.

How to disseminate the results

All translated sheets have been printed on paper sheets as the first user group is in the developing world. When further translating the sheets, local adaptation will have to be made, i.e. language, target group, and education level will have to be taken into account.

A microcomputer data base has already been set up with dBase IV, a programme containing the identification data of the translated sheets. If mass storage capacity will not be a problem, the full sheet may also be a part of this data base. As dBase IV is (semi)relational, it would be simple to establish links, relations, views and joint files with other data bases with complementary information. A data base having several different files and different types of chemical information would be useful [6]. With one single search expression e.g. "Ammonia" one could locate a plain, local language, basic explanation [20], the factories where ammonia is stored in major hazard quantities [21], accidents previously occurring and their prevention recommendations, etc.

All data can also be mounted on CD-ROMs as well as on-line among other already existing CSDSs [11]. This would allow not only searches on limited inverted (indexed) fields but also free-text retrieval, which – when connected with descriptor searching – will clearly improve the retrieval rate and quality.

Those data sheets dealing with major hazard chemicals [21, pp. 51–53] will be part of a microfiche collection on major hazard control information and will be disseminated jointly with a linked data base to be supplied on floppy disks.

Finally, all translated CSDSs will be included as individual records or as a full collection in the CISDOC English and French data bases and disseminated via the three available methods: as printed in the SAFETY AND HEALTH AT WORK - ILO/CIS BULLETIN in five languages, through the on-line hosts (ESA, TELESYSTEM-QUESTEL, MEDLARS-MIC, CCOHS, MAXWELL-PERGAMON, TOXLINE), and the two CD-ROMs [13,14].

Discussion

Machine translation would be easiest if the original text were standardised as much as possible. In the preparation and translation process of CSDSs there are no particular benefits in maintaining a wide variety of different expres-

sions. On the contrary, the simpler the expressions and phrases the less possibilities there are for misinterpretations. This has resulted in the establishment of standardised R- and S-phrases in the labelling and identification systems of the EC [17], and ANSI [22], but a similar expanded system would be highly desirable for the CSDSs as well. The International Programme of Chemical Safety, a joint programme of the WHO, ILO and UNEP, has already taken steps in this direction. Perhaps an extended CIS Thesaurus [12] should also include a few thousand standard phrases with a code linking these to any possible future language. The descriptor system (keyword system) in itself is already a language standardisation method.

The international harmonisation of systems of classification, labelling and identification for the use of hazardous chemicals at work, a process started by an ILO Resolution [23], should consider the standard phrases of chemical safety. A number of key expressions have already been created by existing standards [17,22] and this translation process. This may also include reconsideration and possibly reduction of the number of pictograms and symbols (e.g. symbols for toxic, flammable, oxidising etc.), which are in essence already understandable by most cultures. The background work has already been carried out by the United Nations Committee of Transport of Dangerous Goods [24], although it has only concentrated on acute effects leaving long-term exposure (such as carcinogenicity) questions aside. Standardisation may not only improve understandability and safety but also greatly facilitate the rapidly increasing world trade of chemicals.

The number of synonyms of chemical names is also a serious problem. Although the CAS numbering system does not have any logical or hierarchial background, it is the only system currently available to unambiguously identify chemicals and products. It should not be difficult to use CAS numbers linked to any other classification system and these may be used as an international code in chemical data sheets and data bases. One CAS number may easily have tens of synonyms as shown in the example below.

CHEMICAL NAME SULFURIC ACID
CAS REGISTRY NUMBER 7664-93-9

SYNONYM(S)/TRADE NAME(S)
: ACIDE SULFURIQUE
: ACIDO SOLFORICO
: BOV
: DIPPING ACID
: HYDROGEN SULFATE (DOT)
: MATTING ACID (DOT)
: NORDHAUSEN ACID (DOT)
: OIL OF VITRIOL

: OIL OF VITRIOL (DOT)
: SPENT SULFURIC ACID (DOT)
: SULFURIC ACID (ACGIH, DOT, OSHA)
: SULFURIC ACID, SPENT (DOT)
: SULPHURIC ACID
: SCHWEFELSAEURELOESUNGEN
: VITRIOL BROWN OIL
: VITRIOL, OIL OF (DOT)
: ZWAVELZUUROPLOSSINGEN

The use of microcomputers for machine translation was clearly shown to be feasible and with their increasing capacities there is no doubt that their easy controllability will be a strong supporting argument for micros compared to larger mini or mainframe computers. The hardware itself is not the main limiting factor, but the 640 kb limit of the operating system. New products with OS/2 are already being developed.

Will the system be suitable for other language pairs? Would it be possible to translate any other materials with a similar system? The system may be fairly easily adapted to translate into another anglosaxon language, e.g. into Swedish. When other languages are considered one should, however, bear in mind that:

(1) The source language morphological analyser with its dictionaries must be developed;
(2) The rulebases will have to be redesigned;
(3) The more standard phrases can be used, the easier will be the result and the quality and understandability, i.e. safety increased;
(4) The input data should already exist in a highly classified, organised and magnetic form like that of a data base, present image scanners and character recognition programmes, according to our tests, may not be convenient enough for this purpose
(5) Simple human spelling and typographical mistakes in the input file are creating a lot of interpretation problems.

If relatively large chemical information or other somewhat limited data bases are to be translated, the microcomputer is a feasible instrument. The results of this study are highly encouraging.

Acknowledgements

This work has been supported by the ILO-Finnida Project RAS/87/09/FIN and the Finnish Work Environment Fund.

References

1 M.L. Richardson, Toxic hazard assessment of chemicals. Royal Society of Chemistry, Distribution Centre, Blackhorse Road, Letchworth, Herts. SG6 1HN 1986, 360 pp.

2 J.D. McAteer, D. Lawson, How to use your right to know chemical hazards - A guide to the new hazard communication standards. The Pilgrim Press, 132 W.. 31st Street, New York, NY 10001, 1986, 75 pp.

3 S.G. Hadden, Providing citizens with information about health effects of hazardous chemicals, J. Occup. Med., 31(6) (1989) 528-534.

4 J.W. McLellan, Hazardous substances and the right to know in Canada, Int. Labour Rev., 128(5) (1989) 639-650.

5 TVATM - Identification and labeling of substances hazardous to health, Turvallisuusmääräykset 39, National Board of Labour Protection, Helsinki, 1986, 142 pp. (in Finnish).

6 J. Takala, J. Andersen, Information and training in occupational safety and health - Decision support system for developing countries, Int. Symp. on Health and Environment in Developing Countries, HEDC-86, Extended Abstracts, Institute of Occupational Health, Helsinki, 1986, pp. 102-111.

7 Safety in the use of chemicals at work, Report V(1), International Labour Conference 77th Session 1990, International Labour Office, Geneva, 1989, 18 pp.

8 I.M. Pigott, A General Introduction to the SYSTRAN Machine Translation System, Commission of the European Communities, note, Luxembourg, 1989, 7 pp.

9 National Institute of Occupational Safety and Health, Registry of Toxic Effects of Chemical Substances, RTECS, database in printed, microfiche and CD-ROM form, Cincinnati, OH, 1990 (in [10] and [11]).

10 CHEMBANK, CD-ROM, SilverPlatter Co. Ltd., London, 1990.

11 CCINFOdisc A1, CD-ROM, Chemical Information, Canadian Centre for Occupational Health and Safety, Hamilton, Ont., 1990.

12 CIS Thesaurus, English and French edn., International Occupational Safety and Health Information Centre, ILO, Geneva 100 and 89 pp. 1976, 1989, 9 suppl.

13 CCINFOdisc B, CD-ROM, Occupational Safety and Health Information, Canadian Centre for Occupational Health and Safety, Hamilton, Ont., 1990.

14 OSH-ROM, CD-ROM, SilverPlatter Co. Ltd., London, 1990.

15 L. Kulikov, H. Jäppinen, Automatic translation of a highly constrained language, Proc. 2nd Scandinavian Conf. on Artificial Intelligence, SCAI '89, Tampere, 1989, pp. 904-911.

16 H. Jäppinen and M. Ylilammi, Associative model of morphological analysis: An empirical inquiry, Computat. Linguist. 12 (4) (1986) 257-272.

17 EC Council Directive of 27 June 1967 on the approximation of the laws, regulations and administrative provisions relating to the classification, packaging and labeling of dangerous substances, 67/548/EEC and 16 other Council and Commission Directives, Brussels, 1967-1986.

18 Classification and Labeling of Dangerous Substances Commonly used in Industry, Reference booklet, Labour Department, Government Printer, Hong Kong, 1989, 257 pp. (annexes in Chinese).

19 J. Takala, A.T. Rajah, P. Hasle and S. Angsutham, Dangerous substances - Information, labeling and personal protection, Nat. Inst. for the Improvement of Working Conditions and Environment - Project Technical Report No. 2, Bangkok, 1985, 78 pp.

20 Nat. Inst. for the Improvement of Working Conditions and Environment, Project Final Report, UNDP/ILO, Bangkok, 1986, 52 pp.

21 Major Hazard Control, A practical manual, International Labour Office, Geneva, 1989, 295 pp.

22 American National Standard for Hazardous Industrial Chemicals - Precautionary Labeling, ANSI Z129.1-1988, Chemical Manufacturers Association, Washington, DC, 1988, 65 pp. (annexes).

23 International Labour Conference, 76th session, Resolution concerning harmonisation of systems of classification and labeling for the use of hazardous chemicals at work, Provisional Record No. 23, Geneva, 1989, pp. 28–29.

24 Recommendations on the Transport of Dangerous Goods, 6th rev. edn, United Nations, New York, NY, 1989, 482 pp.