nks
Nordisk kernesikkerhedsforskning
Norrænar kjarnöryggisrannsóknir
Pohjoismainen ydinturvallisuustutkimus
Nordisk kjernesikkerhetsforskning
Nordisk kärnsäkerhetsforskning
Nordic nuclear safety research

NKS-170
ISBN 978-87-7893-235-8

Ling_An: LINGUISTIC ANALYSIS OF NPP INSTRUCTIONS

Fred Karlsson and Laura Salo
Helsingfors universitet, Finland

Björn Wahlström
VTT, Finland

July 2008

Abstract

Projektplanen består av två delprojekt, (1) att undersöka huruvida den tillgängliga datalingvistiska metodiken (SWECG) är tillämplig på säkerhetsföreskrifterna för Forsmarks kärnkraftverk, och (2) att utreda möjligheten att upprätta ett praktiskt fungerande system på basen av den använda metodiken. Svaret på båda frågorna är jakande. ett praktiskt system kan vid behov och om intresse föreligger förverkligas av Lingsfot Ab, Åbo.

Det första textmaterialet erhållet från Forsmark bestod av 392 sidor ur FKA:s Lednings- och kvalitetshandbok (hädanefter förkortad LOK), drygt 77 000 löpord. Texten kom i 18 PDF-filer som konverterades vanlig råtext. Det andra materialet var drygt tio gånger större och representerade fyra nya typer av text: Driftsinstruktioner (D-I), Instruktioner (F-I), Underhålls-instruktioner/El (Und/El) och Underhållsinstruktioner/Mek (Und/-Mek), sammanlagt över 783 000 löpord.

Key words

säkerhetsföreskrifter, kärnkraftverk, språkanalys, kontrollerat språk, SWECG, automatisk språkanalys

NKS-170
ISBN 978-87-7893-235-8

Electronic report, July 2008

The report can be obtained from
NKS Secretariat
NKS-776
P.O. Box 49
DK - 4000 Roskilde, Denmark

Phone +45 4677 4045
Fax +45 4677 4046
www.nks.org
e-mail nks@nks.org

3.6.2008

Ling_An: LINGUISTIC ANALYSIS OF NPP INSTRUCTIONS

Fred Karlsson, Inst för allmän språkvetenskap, PB 9, FI-00014 Helsingfors universitet, Finland
Laura Salo, Inst för allmän språkvetenskap, PB 9, FI-00014 Helsingfors universitet, Finland
Björn Wahlström, VTT, P.O. Box 1000, FI-02044 VTT, Finland

1. Projektets målsättning, partners och finansiering

Projektets målsättning definierades i projektplanen på följande sätt:

> The purpose of the *Ling_an* project is to assess the applicability of linguistic analysis for supporting the writing and maintaining of Nuclear Power Plant (NPP) instructions. There are many reasons to believe that modern computer supported analysis of written texts can improve the semantic consistency and level of formalization of written instructions and procedures. Today instructions and procedures are created manually and are written in natural language according to standards defined by individual nuclear power companies.
>
> Instructions and procedures have achieved a certain degree of formalization through practice and continuous improvements in spite of the fact that they are written using rather traditional methods. However, the formalization is mainly related to the structure of instructions and not to their syntax or contents. A formalized structure with respect to the use of words and syntax is believed to reduce the risk of human error by improving the readability of instructions. Furthermore a successful completion of the project is believed to pave the path towards an analysis of the semantic content of instructions and procedures. If this would be possible, a major step in computer supported analysis of the NPP documentation could be taken.
>
> Even if the long term goals of the project cannot be reached fully, it is still evident that proposed methods have a large potential in supporting document management procedures at the NPPs. Simple applications include, but are not restricted to, consistency checks of used glossary, identification of linguistics ambiguities in instructions, advanced searches for specific linguistic constructs, etc.

Projektets partners och deras respektive uppgifter är följande:

> The *Ling_an* project will analyse the syntax and semantics of selected existing NPP instructions by using research results from computational linguistics developed at **Department of General Linguistics at the University of Helsinki** (UH; ansvarsperson Fred Karlsson). **VTT Industrial Management and Innovation Systems** (VTT; ansvarsperson projektstart till 31.1.2008 Björn Wahlström, ansvarsperson 1.2.2008 till projektslut Heli Talja) will participate in the project by providing a general systems engineering support and **Forsmarks Kraftgrupp AB** (FKA; ansvarsperson Christer Eriksson) will provide the necessary data material (instructions and procedures written in Swedish) to be analysed.

Projektplanen består av två delprojekt, (1) att undersöka huruvida den tillgängliga datalingvistiska metodiken är tillämplig på texterna ifråga, och (2) att utreda möjligheten att upprätta ett praktiskt fungerande system på basen av den använda metodiken.

2. Delprojekt 1: Datalingvistisk metodik

The linguistic part of the project will use existing tools for the automatic analysis of running Swedish text. These tools are especially the parsing system Swedish Constraint Grammar (SWECG) developed at the Department of General Linguistics at the University of Helsinki and presently owned, updated and serviced by Lingsoft, Inc, residing in Turku / Åbo. For any given text, SWECG first normalizes the text and then performs a morphological and shallow syntactic analysis of it.

In practice, this means that (i) misspelt or new words in the text are identified and either corrected or added to the lexicon (in case of specialized important terminology encountered for the first time), (ii) all inflected words are lemmatized, i.e. they are assigned a base form (e.g. "come" and "came" are forms of the verb "come"), (iii) words with several potential lemma representations are disambiguated (e.g. "guide" may be either a noun or a verb, depending on context one or the other lemma is induced to be correct, and (iv) the important phrases of the text are determined (e.g. the three words "big bad wolf" belong together as one phrase with one denotation).

Consider example (1), picked from one of the Forsmark manuals:

(1) "Specifikt gäller att konstruktion skall ske i steg som är avpassade för processens behov, för god spårbarhet och enkel verifiering. För projektledningsaktiviteterna gäller ISO 10006 och PMIs projektmodell som förebild."

Of course, SWECG in its present state has not been tuned to texts with all the special terminology of nuclear power plants. During analysis phase (i), some words will therefore be flagged as unknown to the system, at least "PMIs" in this text excerpt. Probably the lemma "PMI" recurs elsewhere in the texts and it must therefore be added to a special sublexicon containing Forsmark terminology. As more and more texts are fed to SWECG, this sublexicon comes to contain the highly genre-specific words of the nuclear plant documentation. Obviously, this has great intrinsic value for further analysis of the contents of masses of documentation.

During phase (ii), all word-forms are lemmatized, i.e. "Specifikt" is assigned to the adjective lemma "specifik", "avpassade" to the verb lemma "avpassa", "projektledningsaktiviteterna" to "projektledningsaktivitet" but also to "projektledning" (in addition it would be related to the phrases "leda ett projekt" and "aktiviteter för att leda ett projekt"), etc.

After phases (ii) and (iii), a lemmatized and disambiguated word-list has been produced which can be subjected to further analysis, e.g. frequency and content analysis. Phase (iv) is important because it discloses the phrases, i.e. which words go importantly together, such as "PMIs projektmodell".

Further analysis is needed, tailored to the specific nature of these texts. E.g., for the important purpose of making action analysis possible, phases (i, ii) yield the plain verbs of the text ("ske", "avpassa" etc.), but this is not enough because many important actions are expressed by a multiword combination of a verb with a bleak general meaning in combination with a specific noun which expresses the core meaning, e.g. "konstruktion skall ske" must be related to the verb "konstruera". Obviously, this is a non-trivial but indispensable step of analysis.

In conclusion, linguistic analysis of the type sketched makes it possible to:
- spot and correct **misspellings** and **unlicensed vocabulary** (DELIVERABLE 1, sections 5-6 below)
- establish the **core vocabulary** and the **specialized nuclear plant vocabulary** of the Forsmark texts (DELIVERABLE 2, sections 6-7 below),
- establish the **differences between different types of Forsmark texts** (DELIVERABLE 3, section 8)
- **control** that one and the same object and process is consistently referred to by one term only (DELIVERABLE 4, section 9),
- establish a concrete **analyzed text basis** with word lemmata, parts of speech, and phrases indicated on top of which semantic action analysis can build. In this task the focus would be on building a database of Forsmark terminology by iterative application of SWECG to more and more text material. This database can be estimated to contain at least 5,000 items of individual words. Special studies are needed to establish the Forsmark-typical phrases that are in use and that must be split up and related to individual

words (such as "projektlednings-aktiviteterna", "konstruktion skall ske") (DELIVERABLE 5, section 10).

En sammanfattande beskrivning av SWECG:s lämplighet som plattform för NPP instruktioner ingår i analysen av möjligheter till konkret förverkligande, DELIVERABLE 6, avsnitt 10.

Projektet finansierades med 250 000 DKK av Nordisk Kärnkraftssäkerhet (Köpenhamn) samt med 100 000 SEK av Statens Kärnkraftinspektion (Stockholm). Arbetet utfördes vid Institutionen för allmän språkvetenskap, Helsingfors universitet, med professor Fred Karlsson som projektledare samt fil. stud. Nuutti Harmo (1.2.- 30.11. 2007) och fil. stud. Laura Salo (1.11. 2007 – 30.4. 2008) som främsta projektanställda.

För var och en av de i projektplanen omnämnda slutprodukterna 1-5 (DELIVERABLES) levereras till uppdragsgivarna också de datafiler, som innehåller de analyserade materialen beskrivna i respektive avsnitt.

3. Delprojekt 2: Möjligheter till konkret förverkligande

Det andra delprojektet är en analys av möjligheterna att få till stånd ett fungerande system på basen av de nu vunna erfarenheterna (DELIVERABLE 6, avsnitt 10).

Objective
The usability of the approach will be evaluated.

Description
The usability of the approach will be assessed using different criteria such as

- costs and benefits of using linguistic tools in the support of writing and maintaining NPP instructions and procedures,
- possibilities to build a suitably simple “controlled language”, which can reflect the intent and execution of most instructions that are in use at NPPs,
- needs and possibilities for further development efforts to arrive at specified levels of computerised support for document management at NPPs.

Konsulthjälp beträffande delprojekt 2 har givits av VD Juhani Reiman, Lingsoft Ab.

4. Textmaterialet ur kvantitativ synpunkt

Det första textmaterialet erhållet från Forsmark bestod av 392 sidor ur FKA:s *Lednings- och kvalitetshandbok* (hädanefter förkortad LOK), drygt 77 000 löpord. Texten kom i 18 PDF-filer som konverterades vanlig råtext. Det andra materialet var drygt tio gånger större och representerade fyra nya typer av text: Driftsinstruktioner (D-I), Instruktioner (F-I), Underhållsinstruktioner/El (Und/El) och Underhållsinstruktioner/Mek (Und/Mek), sammanlagt över 783 000 löpord.

Konverteringen av materialet och speciellt extraheringen av texten var mycket tidsödande (mera om detta i avsnitt 5), sammanlagt spenderades tre personmånader på detta. Bland annat användes programmen Pdftotext, Antiword samt självgjorda unixskripter.

Tabell 1 framställer hela materialets omfattning över **löpord** (konkreta teckensträngar, alla inkluderade och räknade var för sig) **löpordstyper** (med identiska löpord sammanslagna och hopräknade) samt **lemman** (med sammanhörande löpordstyper hopförda och sammanräknade under grundformer).[1] Alla löpord är med, också de som innehåller (enbart eller delvis) siffror.

materialtyp	löpord	löpordstyper	olika löpordstyper	lemman	olika lemman
LOK	77 624	8 024		6 165	
D-I	344 545	31 691		27 741	
F-I	164 075	19 000		14 605	
Und/El	168 030	17 091		14 809	
Und/Mek	106 445	9 481		7 820	
summa (∑)	860 719	85 287	65 061	71 140	52 928

Tabell 1. Förekomsten av löpord, löpordstyper och lemman i delmaterialen.

När delmaterialen i tabell 1 sammanslås, framkommer det totala antalet olika (unika) löpordstyper (65 061) samt olika lemman (52 928). Dessa tal är förvånansvärt stora. Den viktigaste orsaken är de talrika förekomsterna av **sifferlöpord** (löpord som består enbart av siffror) samt **bokstavssifferlöpord**, dvs ord som består både av siffror och bokstäver. Dessa är mycket typiska för Forsmarktexterna, t.ex. C8-C9, DG310, KAE.104, 1.D0124, EL32.B28.X6, E1.11, och FQ-2007-0199. Detta framställes i tabell 2, som visar fördelningen av sifferlöpord, bokstavssifferlöpord och **bokstavslöpord** (varmed menas ord som består enbart av bokstäver).

1	2	3	4	5	6	7	8	9	10
	siffer-löpords-typer		bokstavs-siffer-löpordstyper		bokstavs-löpords-typer		∑ alla	vanliga lemman	
	n	%	n	%	n	%		n	%
f > 1	4 774	38,91	6 344	51,22	23 950	59,25		17 711	62,64
f = 1	7 495	61,09	6 042	48,78	16 475	40,75		10 562	37,36
∑ typer	12 269	18,86	12 386	19,04	40 406	62,10	65 061		
∑ löpord	105 142	12,22	73 672	8,56	681 728	79,22	860 542		
∑ vanliga lemman								28 273	

Tabell 2. Löpord, löpordsordtyper och lemman i hela materialet (f=frekvens n = antal).

Av löporden är drygt 20% (nästan 180 000) antingen sifferlöpord eller bokstavssifferlöpord. De senare är i allmänhet obekanta för det utnyttjade analyssystemet SWECG (dvs ingår inte i dess interna lexikon) och ger därför upphov till analysproblem. På löpordstypernas nivå är andelen sifferlöpord + bokstavssifferlöpord nästan 38%, vilket understryker det just sagda.

Antalet vanliga lemman (etablerade bokstavsord utan siffror) är över 28 000. Detta antyder en stor lexikal spridning, dvs att språkbruket inte är starkt kontrollerat. Av dessa lemman är

[1] I meningen *Eva, Evas man och Evas barn kom på festen* finns alltså nio löpord, åtta löpordstyper och sju lemman representerade, varav lemmat *Eva* representeras av tre löpord (*Eva, Evas, Evas*) och två löpordstyper (*Eva, Evas*).

1600 verb och nästan 22 000 vanliga substantiv (samt nästan 14 000 bokstavssiffersubstantiv).

5. Problem med textmaterialet

Analysen i detta avsnitt är en del av DELIVERABLE 1. Den andra delen av DELIVERABLE 1 avhandlas i avsnitt 6.

5.1. Metatext

Instruktionerna är verbalt avfattade (inte strukturerade t.ex. enligt XML), vilket innebär att olika typer av text (metatext, instruktioner, listor osv) alla ser ut som samma slags text. Några exempel på metatext: *Sökord (kan även fyllas i nedan)*, *Granskad*, *Godkänd* och *Hänvisningar till följande dok.*

5.2. Sammansättningar

När det gäller sammansättningar förekommer problem i tolkningen. Kombinationer av sammansatta ord, till exempel *led- och manuell drift* och *HT- och LT-reglerventilerna*, orsakar problem speciellt gällande första delen av kombinationen. Ordbetydelsen kräver att också den första delen behåller både dess för- och efterled. Ändå är alla liknande konstruktioner i materialet inte sammansättningar.

Ett annat problem förorsakas av möjliga sammansättningar vars fristående för- eller efterled består av ett forsmarkspecifikt ord. Dessa ord består ofta av både bokstäver och siffror, t.ex. *Återgångssteg R13 – Magnetiseringsdrift*. Det är omöjligt att utan specifik information klassificera *R13* som en del av den första eller andra substantivfrasen.

Det finns forsmarkspecifika förkortningar i sammansatta ord, t.ex. *doktyp* och *författarsign*, som föranleder problem i automatisk morfologisk analys.

5.3. Förkortningar

I materialet förekommer en stor mängd förkortningar som har betydelse endast inom forsmarksramen. *Doktyp, författarsign, F1DD(2), F1D/Arkiv, nöddrän, drän.vent, syst, bif, aut.stängning, 414 P5/P6, FTA 89-288* och *proj.rapp*. Dessa kan ha flera skrivsätt.

5.4. Ellipser

Delvisa ellipser (utelämnanden) förekommer i uttryck som *1 008-1, -2, -4, -5, -6, -7, -9, -11, -16 Utg 3*, som uppenbarligen har nio referenter. I *Dränageventilerna 418 V33/V34, V31/V32, V21/V22* hänvisas till tre referenter och i *Byepass* [sic] *av 415 P305 och P306* till två. Ellipserna kan vara betecknade med både bindestreck och komma eller konjunktion.

5.5. Orddelning med ”-”

Bara i driftsinstruktioner (D-I) finns 373 löpord, som delats med bindestreck (hard hyphen) över två rader. Detta är enkelt att korrigera. Svårare är förekomster där ord har delats så att mellan delarna står en bit av en annan fras:

11.2.2 PC2(5) 411 K852 Öppnar, dränventiler ång-

V527, V536 ledningar

Ord har delats över två rader med bindestreck som ska behålla sitt bindestreck, t.ex. *HT-förvärmare* och *LT-reglerventilerna*. Det finns delade ord som efter konverteringen till råtext inte finns i rader efter varandra (eventuellt hade originalfilformatet varit effektivare att konvertera och under alla omständigheter hade strukturering av dokumenten effektiverat utdumpningen av den egentliga relevanta texten):

KAPACI-
1
749-9
TETSPROV
749

5.6. Bristande systematik

Ellipser förekommer i två versioner, bindestreck och komma / konjunktion. Förkortningarnas form varierar: *avt* vs. *avt.,* *enl* vs. *enl.;* användningen av mellanslag med snedlinje: *uppgång/nedgång* vs. *Varvtal / effektbörvärde*; *dränvent.* vs *drän.vent* och å andra sidan *drän.vent* vs. *dränageventiler*; beteckningen av specialtecken som gradtecken *150(c* vs. *140oC* vs. *100 c*.

5.7. Felskrivningar

Det finns en betydlig mängd felskrivningar i materialet: bortfallna mellanslag, t.ex. *pumpUtgått, bakventavt, övervakningVälj, utlösta.Välj*; andra slagfel: *reglerrar, inkopplinge, falelt, Partell;* användningen av skiljetecken: *uppgång/nedgång”effektbörvärde.*

5.8. Enheter bestående av siffror

Begreppet ”ord” i forsmarktexter är komplext. Förutom att ellipser, orddelning, felskrivningar och brist på systematik i beteckningar ökar risken att ett och samma ord klassificeras som olika lemman, är ett ytterligare problem det att strängar av siffror ofta bör uppfattas som en enhet motsvarande ett ord (eller en del av ett ord): *Beskrivande del flyttad från 1 008-17 till 1 008-8* och *återgångskontroll 1Z3.22, fortsätt i DI 1 008-13 mellansteg 2Z2.* Många av dessa enheter innehåller mellanslag som en integrerande del.

5.9. Bristande användning av stora skiljetecken

Ett stort problem för de nuvarande analysverktygen föranleds av att texterna inte utgörs av traditionella ”meningar” som avslutas med stora skiljetecken (punkt, frågetecken, utropstecken). Analysverktygen utgår från att de matas med traditionella meningar.

5.10. Ostrukturerade dokument

Dokumenten (i formen vi fick dem) är vanliga pdf-filer, dvs. ostrukturerade, utan att olika typer av text skulle ha särskiljts (rubriker, instruktioner, listor, metatext, uppdateringsuppgifter etc.). Detta gör att det viktigaste (instruktionerna) inte kan identifieras automatiskt.

Ytterligare exempel på problemtyperna 5.1 – 5-10 ges i filen *textproblem.doc*, som också är en del av DELIVERABLE 1.

6. Automatisk identifiering av felskrivna ord, otillåtna ord samt nya ord som skall tillfogas lexikonet

Analysen i detta avsnitt är en del av DELIVERABLE 1 och hela DELIVERABLE 2. En annan del av DELIVERABLE 1 avhandlas i avsnitt 5.

SWECG accepterar som tillåtna exakt de ord (inklusive böjningsformer och sammansättningar) som finns upptagna i SWECG:s baslexikon, i praktiken för närvarande hela svenskans grundläggande ordförråd enligt Svenska Akademiens ordlista, kompletterad med namn och annan specialvokabulär. Om SWECG konfronteras med ord (teckensträngar) som inte är legala former (inklusive böjningsformer och sammansättningar) av orden i lexikonet, flaggas dessa ord som okända.

I praktiken kan dessa ord vara av två typer, antingen (i) felskrivna eller (ii) rättskrivna nya ord, som bör läggas till baslexikonet för att senare (inklusive böjningsformer och sammansättningar) godkännas som legala.

SWECG behandlar alla strängar bestående enbart av siffror som legala (sifferlöpord). Felskrivna sifferlöpord kan alltså inte igenkännas på automatisk väg av SWECG.

Bokstavssifferlöporden (bestående av bokstäver+siffror) är mycket typiska för Forsmarktexterna. Materialet innehåller 73 762 bokstavssifferlöpord fördelade på 12 386 typer (tabell 2), nästan alla okända för SWECG (som känner igen bara typer börjande med siffror efterföljda av legala bokstavssammansättningar som *326-ledning*, *324-massa*, *321-varv_tals_styrning*). De tio vanligaste bokstavssifferlöporden är i tabell 3.

f	ord
1355	pa4
800	pb1
671	l1
632	h1
613	pa3
464	v4
431	pb7
417	p2
379	pa2
372	p1

Tabell 3. De tio vanligaste bokstavssifferlöpordstyperna (f = frekvens).

Bokstavssifferlöporden upptas i filen *bokstsiff_lopord.xls*. Många är frekventa (tabell 4).

f	n
1000	1
500	5
100	86
50	243
10	1 157
5	2 166
1	6 344
0	12 385

Tabell 4. Antalet bokstavssifferlöpordstyper (*n*) med frekvens större än *f*.

Av de 6000 bokstavssifferlöpordstyperna med frekvensen 1 är många felskrivna. Av dem som förekommer mer än en gång är flera tusen namn på processer, delar, dokument osv, som i ett verkligt system borde läggas till lexikonet. Också vissa av orden med frekvensen 1 bör läggas till lexikonet. Vilka detta gäller i detalj kan endast avgöras av en expert.

Det finns 8184 (nya och felskrivna) bokstavslöpordstyper som SWECG inte känner igen (i filen *bokst_lopord_inte_SWECG.xls*), textfrekvensen är 54 379. De tio vanligaste är i tabell 5.

f	ord	full form
2432	pos	position
2024	sign	signerad, signatur
1750	anm	anmärkning
1679	plac	placering
1323	sub	
1223	enl	enligt
683	fka	Forsmarks Kraftgrupp AB
570	aprm	
485	rä	
437	utg	utgåva

Tabell 5. De tio vanligaste icke-identifierade bokstavslöpordstyperna (f = frekvens).

Flera tusen bör inkluderas i lexikonet, t.ex. *inläckage*, *mava-pumpar*, *nedstyrning*. Vilka detta gäller i detalj kan endast avgöras av en expert. Drygt 4100 av de icke-identifierade bokstavslöpordstyperna har frekvensen 1 och många av dessa är felskrivningar. Tabell 6 återger de icke-identifierade bokstavslöpordstypernas frekvensfördelning.

f	n
500	8
100	50
50	132
10	741
5	1 371
1	4 123
0	8 212

Tabell 6. Antalet icke-identifierade bokstavslöpordstyper (*n*) med frekvens större än *f*.

Ifall man vill införa en strikt regim för vilka ord som är förbjudna att använda och vilka därför bör ersättas med andra uttryck, kan de förbjudna orden flaggas i lexikonet för SWECG på så sätt, att varje gång en form av ett sådant ord påträffas i en (gammal eller ny) Forsmarktext, meddelas i programmets analys vilket det korrekta uttrycket borde vara.

Filen *nya_i_SWECG.xls* innehåller drygt 6300 vanliga (med bokstäver skrivna) lemman som borde fogas till lexikonet.

7. Det centrala ordförrådet och det kärnkraftsspecifika ordförrådet

I detta avsnitt presenteras den andra delen av DELIVERABLE 2. Den första delen återfinns i avsnitt 6 (sifferbokstavslöporden, tabell 4).

För att kunna avgöra vilka ord som skall anses höra till svenskans grundläggande ordförråd, vilka åter som är typiska för Forsmarktexterna (”kärnkraftsspecifika”), behövs ett jämförelse-material. Som sådant utnyttjas här PAROLE-materialet, över 19 miljoner löpord svensk text hopsamlad i slutet på 1990-talet (se http://spraakbanken.gu.se/parole/). Detta materials sammansättning återges i tabell 7.

Romaner	Bonniers Grafiska Industrier	1976—1981	4.4 milj. löpord
Dagstidningar	Dagens Nyheter Svenska Dagbladet Göteborgs-Posten Arbetet	1976—1997	13.6 milj. löpord
Tidskrifter	Forskning och Framsteg	1995—1996	0.4 milj. löpord
Övrigt	Webb-texter	1997	1 milj. löpord

Tabell 7. PAROLE-materialets sammansättning (19+ miljoner löpord).

Antalet löpordstyper i PAROLE är 573 547. Detta är materialet ”hela PAROLE” i tabell 8 nedan. Dessutom ingår i tabell 8 två mindre jämförelsematerial extraherade ur PAROLE, de 2000 vanligast löpordstyperna (P2000) och de 5000 vanligaste löpordstyperna (P5000). Med dessa material som antas representera typisk svenska jämförs bokstavslöpordstyperna och lemman i Forsmarkstexterna (tabell 8).

	hela PAROLE		P2000		P5000	
Forsmark	i bägge	inte i. PAROLE	i bägge	inte i P2000	i bägge	inte i P5000
bokstavslöpordstyper	16 990	23 416	1 471	38 935	2 867	37 539
bokstavslemman	6 952	21 295	897	27 350	1 643	26 604

Tabell 8. Forsmarkorden jämförda med hela PAROLE, P2000 och P5000.

Det finns alltså över 21 000 bokstavslemman som inte förekommer överhuvudtaget i PAROLE. Härtill bör läggas en stor del av de över 12 000 bokstavssifferlöpordstyperna presenterade i tabell 4. Dessa data tyder på att det kärnkraftsspecifika lexikonets storlek överstiger 30 000 lemman.

Materialet underliggande de ovan nämnda resultaten bifogas i sex filer:

bokst_lemman_inte_i_hela_PAROLE.xls
bokst_lemman_inte_i_P5000.xls
bokst_lemman_inte_i_P2000.xls
bokst_typer_inte_i_hela_PAROLE.xls
bokst_typer_inte_i_P5000.xls
bokst_typer_inte_i_P2000.xls

I filen *verblemman.xls* finns texternas verb, som är instruktionernas ryggrad. Här listas 1599 verblemman, vars textfrekvens är 84 868. De existerande verblöporden återfinns i filen *verblopord.xls*, till antalet 3753. Vanliga substantivlemman (skrivna med bokstäver) finns i filen *bokst_subst_lemman_2.xls*, till antalet 21 357 (av vilka över 12 700 är sammansättningar markerade med '_'), representerande 282 284 löpord.

8. Ordvariation mellan olika typer av Forsmarktexter

Detta är DELIVERABLE 3, som redogör för variationen mellan orden i olika typer av F-texter.

Ling_An-projektet har analyserat fem olika typer av Forsmarktexter, presenterade i början av avsnitt 4. Tabell 9 framställer i vilken mån dessa fem texttyper är homogena. Kolumnerna för lemman och typer avser grundformer och löpordstyper. Kolumnerna "alla" förtecknar respektive alla lemman och alla löpordstyper, kolumnerna "icke-p2000" tar upp endast de grundformer och löpordstyper som inte förekommer i jämförelsematerialet PAROLE med 19+ miljoner löpordstyper svensk text (genom att lämna bort de 2000 vanligaste löpordstyperna maximeras det kärnkraftstypiska ordförrådet). Dessutom presenteras data kolumnvis så, att de ges både med och utan sifferlöpord.

Spridningen för data presenteras i tabell 9 horisontellt för lemman och typer som förekommer i respektive 1, 2, 3, 4 och 5 texttyper.

Sammandrag	**lemman**				**typer**			
	alla		icke-p2000		alla		icke-p2000	
	med siff	utan siff	med siff	utan siff	med siff	utan siff	med siff	utan siff
5	884	802	490	**408**	987	866	556	435
4	1010	849	811	**650**	1429	1161	1160	892
3	1771	1473	1627	**1329**	2380	2086	2177	1883
2	5361	4717	5180	**4536**	7212	6539	6924	6251
0 + 1	46645	35213	46458	**35026**	53072	42140	52792	41860

Tabell 9. Texttypernas homogenitet.

Tabell 9 utvisar att spridningen mellan texterna är stor. Mindre än 1000 lemman och löpord förekommer i alla fem texttyper, och av dessa är majoriteten ord som inte hör till det allmänna svenska ordförrådet (och därför finns inte med bland de 2000 vanligaste orden i PAROLE). Den stora majoriteten av lemman och löpord förekommer endast i en texttyp; texterna handlar alltså mestadels om specifika aspekter (objekt, processer) som inte återkommer i de andra texttyperna. Det är uppenbart att det specifika kärnkraftsordförrådet omfattar uppemot 40 000 ord, som antyds av den halvfeta kolumnen i tabell 9 (lemman som inte förekommer i PAROLE och inte innehåller sifferord). Jfr också analysen i avsnitt 7. Tabell 10 ger en mera detaljerad bild av de samma relationerna, med specifik jämförelse av

de olika texttyperna (den översta avdelningen ”0 texttyper” är en liten anomali som kan lämnas obeaktad).

	lemman				**typer**			
	alla		icke-p2000		alla		icke-p2000	
	med siff	utan siff	med siff	utan siff	med siff	utan siff	med siff	utan siff
0 texttyper	24	24	24	24	24	24	24	24
1 texttyp								
D-I	21579	13995	21557	13973	23403	15135	23369	15101
F-I	8657	7630	8549	7522	11086	10620	10919	10453
EL	9644	8006	9622	7984	10386	8765	10353	8731
MEK	4197	3056	4184	3043	4853	4306	4832	4285
LOK	2544	2502	2522	2480	3320	3290	3295	3266
summa	**46621**	**35189**	**46434**	**35002**	**53048**	**42116**	**52768**	**41836**
jmf.	**46645**	**35213**	**46458**	**35026**	**53072**	**42140**	**52792**	**41860**
2 texttyper								
lok-mek	40	39	38	37	49	46	46	43
lok-fi	1522	1507	1448	1433	2039	2025	1922	1908
lok-el	37	36	34	33	49	49	44	44
lok-di	107	90	105	88	155	124	152	121
fi-mek	303	253	290	240	380	351	361	332
fi-el	355	302	324	271	452	428	413	389
fi-di	631	547	600	516	935	857	877	799
di-mek	485	420	475	410	641	586	631	576
di-el	1678	1370	1668	1360	2250	1861	2224	1835
mek-el	203	153	198	148	262	212	254	204
	5361	**4717**	**5180**	**4536**	**7212**	**6539**	**6924**	**6251**
3 texttyper								
fi-el-mek	117	98	105	86	139	116	122	99
di-el-mek	450	289	434	273	571	435	544	408
fi-di-mek	197	179	180	162	301	259	278	236
fi-di-el	439	360	398	319	590	518	530	458
di-lok-el	42	30	39	27	56	41	54	39
el-lok-mek	5	3	5	3	5	5	5	5
di-lok-mek	10	8	9	7	27	26	27	26
fi-lok-mek	140	139	127	126	166	163	151	148
fi-lok-el	101	101	87	87	122	121	105	104
fi-lok-di	270	266	243	239	403	402	361	360
	1771	**1473**	**1627**	**1329**	**2380**	**2086**	**2177**	**1883**
4 texttyper								
fi-el-mek-lok	45	42	38	35	61	58	46	43
di-el-mek-lok	26	15	23	12	34	21	33	20
fi-di-el-lok	221	214	181	174	329	318	267	256
fi-mek-di-el	550	411	427	288	786	546	627	387
fi-di-mek-lok	168	167	142	141	219	218	187	186
	1010	**849**	**811**	**650**	**1429**	**1161**	**1160**	**892**
5 texttyper	**884**	**802**	**490**	**408**	**987**	**866**	**556**	**435**

Tabell 10. Ordvariation i de olika typerna av Forsmarktexter.

9. Kontroll av terminologisk systematik

DELIVERABLE 4 skulle utgöras av en kontroll av att samma objekt och process systematiskt i Forsmarktexterna hänvisas till med samma term. Ling_An-projektets resurser har dock inte räckt till för att genomföra detta arbetsmoment.

En orsak är, att det kärnkraftsspecifika ordförrådet är mycket mera omfattande än de ursprungliga estimaten utgick ifrån (40 000, inte 5 000 specifika kärnkraftsord). Den andra orsaken är de tidsödande problemen med att få originaltexterna utdumpade för analys. Den tredje (och viktigaste) orsaken är, att arbete med terminologisk systematik obetingat förutsätter fackkunskap som Ling_An-projektets lingvistiska deltagare inte besitter.

10. Databas för analyserade Forsmarktexter

DELIVERABLE 5 utgörs av en lista på viktiga verb samt tre textdatabaser innehållande Forsmarktexterna i språkligt analyserad form, med speciell tonvikt på huvudverb, som utgör kärnan i instruktionerna. Björn Wahlström valde ut de 161 viktigaste verben, vilka förtecknas i filen *viktiga_verblemman.xls*. En typisk instruktion uttrycker en handling (aktion, process), som utförs av någon eller något och riktar sig mot något, eventuellt under angivande av villkor, förutsättningar, konsekvenser etc. Det centrala ordet i en instruktion är huvudverbet. Alla (huvud)verb har identifierats och extraherats i sina satssammanhang, så att man lätt kan se ”vad/vem som gör något åt något”. Så här ser databaserna ut (nyckelordet i exemplet är verbet *smörja*):

					smörj	gängan	Med	grafit	disper-gerad	i
					smörj	gängorna	På	sonden	,	muttern
					smörj	in	lastbä-rarens	bakre	Lagerhalvor	med
					smörj	mutt-rarnas	Och	skruv-hålens	Gängor	med
					smörj	rörliga	Delar	med	Godkänt	smörj-medel
					smörj	skruv-hålens	gängor	enligt	Bilaga	2
					smörj	ventil-över-delens	gängor	med	Silikonolja	fore
och	dess	tätningar	är	oska-dade	**smörjs**	cylindern	Och	monteras	Ihop	.
	luckan	öppnas	,	gångjärn	**smörjs**	och	tätnings-listerna	kontrol-leras	,	
				tätnings-lister	**smörjs**	.				
rengörs	varefter	detal-jerna	ocular-kontrol-leras	och	**smörjs**	varefter	cylindern	monteras	Ihop	.

Förekomsterna (sammanlagt cirka 11 500) av de 161 viktigaste verben med tillhörande analyser finns i följande filer:

verbdatabas_viktiga_lemman.xls	verbdatabas i lemmaform
verbdatabas_viktiga_typer.xls	verbdatabas i löpordstypform (innehåller också lemman)
verbdatabas_viktiga_med_SWECG_infon.xls	verbdatabas som innehåller all SWECG-information

11. Behövliga åtgärder om ett verkligt system baserat på SWECG implementeras

Detta avsnitt är DELIVERABLE 6 och skall innehålla en analys av möjligheterna att få till stånd ett fungerande system på basen av de nu vunna erfarenheterna. Punkterna 1-10 nedan är arbetsmoment som borde utföras enligt det mest omfattande konceptet för kontrollerade språk.

Punkterna 1-10 är inte stafflerade enligt kronologi eller lätthetsgrad. Det första steget mot ett mera kontrollerat system vore förverkligande av punkt 8, att överföra dokumentationen i en form som är lättare att behandla.

1. *Uppdatering av lexikonet i SWECG med Forsmarkspecifika ord*. Som framgått av texten, bör ett tilläggslexikon upprättas med Forsmarkspecifik kärnkraftsterminologi, speciellt bokstavssifferord, andra förkortningar samt specifika substantiv och verb, tillsammans uppgående till cirka 30 000 ord. Dessa ord är redan identifierade av Ling_An-projektet och bifogas som skilda filer. Att lägga dessa ord till SWECG:s existerande lexikon och förse dem med behövlig information om böjningsklass och andra språkliga drag kräver några månaders arbete. Efter att dessa ord integrerats i SWECG, känner programmet igen alla legala former av dessa ord, också de som ingår i sammansatta ord. Efter detta arbetsmoment kan SWECG i dess nya version användas som språkkontrollverktyg för (gamla och nya) Forsmarktexter.
2. *Korrigering av direkta språkfel i originaltexterna* (delvis med existerande språkkontrollprogram, delvis manuellt). Dessa fel är till största delen triviala: slagfel och slarvfel. De nedsätter analysprogrammets effekt och borde korrigeras. Skrivfelen har också identifierats av Ling_An och kan i princip till stor del korrigeras automatiskt.
3. *Annan standardisering av originaltexterna*, t.ex. användningen av skiljetecken, eliminering av delningstecken i radslut, minimering av ellipser i sammansatta ord. problemen med skiljetecken som saknas i meningsslut är kännbart, för SWECG förutsätter inmatning av välformade meningar. Om meningsavslutande skiljetecken saknas, går programmet lättare vilse och ger felaktiga analyser.
4. *Terminologisk minimering och standardisering* (dvs. kontroll att samma objekt, process etc. systematiskt hänvisas till med en och samma term). Som ovan noterats kräver detta arbetsinsats av experter på kärnkraftsteknologins termer och språkanvändning.
5. *Begrepps- och extensionsdefiniering* av icke-självklara ord (speciellt

sifferbokstavsord som är Forsmarkspecifika). Kräver expertinsatser.

6. *Standardisering av menings- och satsstrukturen* (kontrollerad syntax). Bör noggrant övervägas om sådan behövs. I nuläget saknas restriktioner.
7. *Regimenterad införsel av nya ord* (som genast införs också i SWECG). Kräver insats av terminologisk expertis.
8. *Strukturering av dokumenten* med t.ex. XML (eXtended Markup Language). Kräver insats av expert på texthantering.
9. *Förnyad automatisk analys av menings- och satsstrukturen* efter det att (i) lexikonet uppdaterats, (ii) dokumenten strukturerats, och (iii) systematisk användning av skiljetecken införts.
10. *Eventuellt semantisk aktionsanalys* på basen av steg (9).

I nära samarbete med Lingsoft Ab, Åbo, har Ling_An-projektet tagit fram följande mera detaljerade underlag för eventuell praktisk implementering (se också den bifogade filen *ProofingSolution_Forsmark_v4.ppt,* som innehåller Lingsofts presentation vid avslutningsmötet i Helsingfors den 23 april 2008).

Allmänt

Projektrapporten kan användas som utgångspunkt för utvecklandet av Forsmarkspecifika språkteknologiska redskap, vars kvalitet och användbarhet i hög grad är beroende av hur väl analysverktygen träffar det Forsmarkspecifika textinnehållet. Projektet exemplifierar en realistisk samarbetsmodell: analysverktygen vidareutvecklas på basen av forskning och utredning varefter lösningarna skräddarsys åt kunden med hjälp av företagsspecifika produkter och lösningar. De utvecklingsmål som ovan skisserats i början av detta kapitel förefaller vara möjliga att förverkliga, somliga relativt snabbt, andra på sikt.

Målsättning

Hur gå vidare med praktiska redskap för textuella arbetsprocesser? Vilka är målsättningarna, nyttorna och riskerna, milstolparna och skedena samt kostnaderna?

Ett allmänt mål är konsekvent framskridning i modulära steg med väl utstakad utvecklingsstig. Lösningen för med sig nytta i alla skeden, början från språkkontrollverktyg för dokumentering i hela organisationen till mera genomgripande interventioner i processerna. Vad gäller t.ex. språkkontrollverktyg är ett bra slutresultat synnerligen sannolikt, eftersom de Forsmarkspecifika lösningarna kan konstrueras med befintlig och välprövad teknologi. Utvecklingsansträngningarna kan då uttryckligen fokuseras på innehållet.

I det stora hela är målsättningen att överföra det språkkunnande som behövs i textuella arbetsprocesser från terminologier, stilguider, arbetsdirektiv o.d. till en integrerad del av de redskap som används i processen. Med andra ord,att använda integrerade språkkontrollverktyg för implementering av bästa tänkbara praktiker för kontrollerat språkbruk.

Projektets faser allmänt

Det är synnerligen viktigt att göra en ordentlig definiering av kundbehoven. I det nu föreliggande fallet har grundarbetet gjorts som dokumenterats i föreliggande rapport. Allmänt taget kan man se följande faser för projekt med målsättning att utveckla kund- eller ämnesområdesspecifika språkteknologiska resurser och redskap:

1. Lingvistiskt arbete, i detta fall exemplifierat av Ling_An-projektet. Denna fas kan bestå av:

a. första skedets korpusarbete på basen av föreliggande textmaterial, men utan att det nödvändigtvis kräver intensiv kundkonsultation

b. språkkontrollverktygen byggs i nära samarbete med kunden, speciellt när man går djupare i fackkunskap gällande språkbruket, exempelvis validerade, rekommenderade och förbjudna termer och fraser.

2. Tillämpning av det lingvistiska arbetet i språkredskap såsom rättstavningskontroll, grammatik- och stilkontroll samt sökindexering.

3. Den funktionella arkitekturen, dvs. vilka funktionaliteter lösningen kräver (bl.a. terminologier, administration och roller).

4. Tillämpning av den funktionella arkitekturen i organisationens processer och applikationsmiljön (bl.a. textbehandlingsprogram, dokumentationssystem och programmeringsmiljöer).

Definition av projektets faser

Fas 1: vidareutvecklad Forsmark-CG (”Constraint Grammar” = restriktionsgrammatik, dvs specifikt Forsmark-inriktad automatisk språklig analys av ord och meningar)

Lexikalisering (tillägg till det redan existerande elektroniska allmänna svenska lexikonet) av det material som extraherats som ett resultat av projektet (DELIVERABLE 6) i morfologianalysatorn SWETWOL för restriktionsgrammatikparsern SWECG för att uppnå förbättrad analys efter att analysatorn känner igen det kärnkraft- och Forsmark-specifika ordförrådet. Kompilering till exekutivt program samt testning. Utgör grundval för de redskap som utvecklas i kommande faser.

- Omfattning: 3 personmånader

Fas 2a: produktiva redskap på basen av Forsmark-CG

I denna fas skräddarsys språkteknologiska redskap som baseras på en berikad Forsmark-CG, exempelvis rättstavning, grammatik- och stilkontroll, kontrollerat/strukturerat språk, kvalitetskontroll, nyckelord och sökindexering, text mining (extrahering, klassificering, klusterering), osv.

Språkkontroll

Man kan skilja mellan allmänspråk, ämnesområdesspråk och organisationsspråk. Nyttan av skräddarsydd rättstavning är desto större ju djupare man skräddarsyr. Studier har visat att språkkontroll av samma typ av funktionalitet och användargränssnitt som i MS Word ger bästa resultat.

Samma kontrollprogram är lätt att integrera också i andra redskap som används i textuella processer, dvs. rättstavning för allmänspråk kan utan ytterligare lingvistisk utveckling integreras i andra editorer, där rättstavning inte är tillgänglig. Utvecklingsstigen är uppenbar: först allmänspråk, sedan ämnesområde och slutligen organisation (Forsmark-språkkontollprogrammet).

Språkkontrollen bör genomsyra hela textframställningsprocessen, så att fel uppdagas med hjälp av språkkontrollen så fort som möjligt, men så att kontrollen signalerar felen i senare skeden av processen. Skräddarsydd kontroll kan utvidgas till andra språk, såsom finska, danska, engelska osv.

Utvecklandet av språkkontroll sammanhänger med terminologihantering. Lingsoft Ab har erfarenhet och redskap för en väl sammanhållen process, där termarbetet är sammanflätat med utvecklandet av språkkontrollprogrammen. Med välutarbetad arbetsfördelning är det fråga om en investering som mycket snabbt ger utdelning, men vars effekter sträcker sig långt i framtiden.

- Forsmark-rättstavning, version 1: kostnad: 3 personmånaders arbete plus licenspris för kontrollprogrammet. Dessutom eventuell integrering i miljöer som ännu inte är stödda.
- Fördelar:

    - inbesparingar eftersom (en del av) språkkorrekturen automatiseras (inbesparing av 5-10 personår i expertorganisation med 100 användare)
    - förbättrad kvalitet

- Referenser: Svefix (= språkkontroll av allmänt svenskt högspråk), EU-kontroll, Microsoft-kontroll för lokaliseringsproffs
- Innehållet skräddarsys tillsammans med kundens experter. Lingsoft står för det tekniska datalingvistiska förverkligandet.

Sökning

Samma analysredskap kan också användas för att berika sökning. Ett välbeprövat sätt är indexering av lemman. Sökordet ”hand” hittar då dokument där ordet har formen ”händernas”.

- fördel: inbesparningar som resultat av effektivare sökning
- kostnad: integrering i kundens användarmiljö. Licens enligt antalet användare.

Utöver dehär kan man skissera andra redskap och möjligheter, som dryftats i rapportens kapitel 11

Fas 2b

Implementering av redskapen i konkreta processer, applikationer och miljöer

Bihang 1: bifogade datafiler med beskrivningar på innehållet

(Dessa filer distribueras till uppdragsgivaren Forsmarks Kraftgrupp AB vid projektslut i maj 2008.)

FILNAMN	DELIVERABLE
textproblem.doc (exempel på problem i originaltexterna)	1
bokstsiff_lopord.xls (löpordstyper bestående av bokstäver och siffror)	1
bokst_lopord_inte_SWECG.xls (löpordstyper bestående enbart av bokstäver)	1
bokstsiff_lopord.xls (löpordstyper bestående av bokstäver och siffror)	2
bokst_lemman_inte_i_hela_PAROLE.xls (bokstavslemman som inte förekommer i hela PAROLE-materialet)	2
bokst_lemman_inte_i_P5000.xls (bokstavslemman som inte förekommer bland PAROLE-materialets 5000 vanligaste ord)	2
bokst_lemman_inte_i_P2000.xls (bokstavslemman som inte förekommer bland PAROLE-materialets 2000 vanligaste ord)	2
bokst_typer_inte_i_hela_PAROLE.xls (bokstavslöpordstyper som inte förekommer i hela PAROLE-materialet)	2
bokst_typer_inte_i_P5000.xls (bokstavslöpordstyper som inte förekommer bland PAROLE-materialets 5000 vanligaste ord)	2
bokst_typer_inte_i_P2000.xls (bokstavslöpordstyper som inte förekommer bland PAROLE-materialets 2000 vanligaste ord)	2
bokst_subst_lemman_2.xls (de 21 357 vanligaste bokstavssubstantivlemmata)	2
verblemman.xls (alla verblemman i de fem texttyperna)	2
verblopord.xls (alla verblöpord i de fem texttyperna)	2
nya_i_SWECG.xls (nya bokstavsord att tillfoga SWECG-lexikonet)	2
(inga skilda filer ingår)	3
(inga skilda filer ingår)	4
viktiga_verblemman.xls (de 161 viktigaste verben, utvalda av Björn Wahlström)	5
verbdatabas_viktiga_lemman.xls (verbdatabas i lemmaform)	5

verbdatabas_viktiga_typer.xls 5
(verbdatabas i löpordstypform (innehåller också lemman)

verbdatabas_viktiga_med_SWECG_infon.xls 5
(verbdatabas som innehåller all SWECG-information)

ProofingSolution_Forsmark_v4.ppt 6
(Lingsoft Ab:s detaljerade förslag till implementering)

Bihang 2: deltagare vid avslutningsmötet i Helsingfors 23 april 2008

Andersson, Sven-Åke	Vattenfall
Eriksson, Christer	Forsmark
Isaksson, Patrik	NKS
Johansson, Peter	Vattenfall
Karlsson, Fred	Inst för allmän språkvetenskap, HU
Karlsson, Monica	Vattenfall
Martini, Peter	Vattenfall
Reiman, Juhani	Lingsoft
Salo, Laura	Inst för allmän språkvetenskap, HU
Selänniemi, Juhani	Lingsoft
Talja, Heli	VTT
Timonen, Juhani	Swot Consulting Finland Ltd
Wahlström Björn	VTT
Åhnberg, Mikael	Forsmark

Bibliographic Data Sheet **NKS-170**

Title	Ling_An: LINGUISTIC ANALYSIS OF NPP INSTRUCTIONS
Author(s)	Fred Karlsson[1)], Laura Salo[1)] and Björn Wahlström[2)]
Affiliation(s)	[1)]Institutionen för allmän språkvetenskap, Helsingfors universitet, Finland [2)]VTT, Finland
ISBN	978-87-7893-235-8
Date	July 2008
Project	NKS-R / LingAn
No. of pages	19
No. of tables	10
No. of illustrations	0
No. of references	0

Abstract

Projektplanen består av två delprojekt, (1) att undersöka huruvida den tillgängliga datalingvistiska metodiken (SWECG) är tillämplig på säkerhetsföreskrifterna för Forsmarks kärnkraftverk, och (2) att utreda möjligheten att upprätta ett praktiskt fungerande system på basen av den använda metodiken. Svaret på båda frågorna är jakande. ett praktiskt system kan vid behov och om intresse föreligger förverkligas av Lingsfot Ab, Åbo.

Det första textmaterialet erhållet från Forsmark bestod av 392 sidor ur FKA:s Lednings- och kvalitetshandbok (hädanefter förkortad LOK), drygt 77 000 löpord. Texten kom i 18 PDF-filer som konverterades vanlig råtext. Det andra materialet var drygt tio gånger större och representerade fyra nya typer av text: Driftsinstruktioner (D-I), Instruktioner (F-I), Underhålls-instruktioner/El (Und/El) och Underhållsinstruktioner/Mek (Und/-Mek), sammanlagt över 783 000 löpord

Key words säkerhetsföreskrifter, kärnkraftverk, språkanalys, kontrollerat språk, SWECG, automatisk språkanalys

Available on request from the NKS Secretariat, P.O.Box 49, DK-4000 Roskilde, Denmark.
Phone (+45) 4677 4045, fax (+45) 4677 4046, e-mail nks@nks.org, www.nks.org