



UPPSALA
UNIVERSITET

Institutionen för lingvistik och filologi
Språkteknologiprogrammet
Examensarbete i datorlingvistik
24 januari 2008

Utvärdering av manuell och automatisk termextraktion

Mirza Škornja

Handledare:

Beáta Megyesi, Uppsala Universitet

Anna Sågvall Hein, Uppsala Universitet

Henrik Nilsson, Tekniska Nomenklaturcentralen (TNC)

Abstrakt

Detta examensarbete är en utvärdering av manuell och automatisk termextraktion. Utvärderingen består av en jämförelse av att ta ut termerna manuellt och att göra det automatiskt för att sedan sammanställa resultatet av dem båda. Resultatet mäts med precision och täckning, beräknat i procent. För att uppnå ett manuellt resultat med höga procent på precision och täckning stöter man på problem som inte kan lösas av en individ. Metoden är inkonsekvent, tids- och resurskrävande. Det automatiska resultatet uppnår som högst 33,3 % på täckning och 20,7 % på precision, men detta är resultatet av tre olika inställningar av programmet Trados MultiTerm Extract, som har till uppgift att ta ut termer ur en text maskinellt. Dock finns det stora fördelar med verktyget som gör att det kan vara något värt att satsa på i framtiden.

Abstract

This thesis is an evaluation of manual and automatic term extraction. The evaluation consists of a comparison between manually extracting terms and doing it automatically. The term candidate lists obtained is evaluated by precision and recall. To achieve a manual result with high percentage on precision and recall, you will come across problems that cannot be solved on your own. The method is inconsistent and demands loads of time and resources. The automated result reaches up as highest 33,3 % when it comes to recall and 20,7 % when it comes to precision. That is the result of using three different settings of the software Trados MultiTerm Extract, which extracts terms from a text automatically. Nevertheless the tool possesses great advantages which can make it worth investing in for the future.

Innehåll

| | |
|--|-----------|
| Abstrakt | 2 |
| Tack | 4 |
| 1 Introduktion | 5 |
| 1.1 Syfte | 5 |
| 1.2 Struktur | 6 |
| 2 Terminologi och terminologiarbete | 7 |
| 2.1 Teorin | 7 |
| 2.2 Terminologiarbetet | 9 |
| 2.3 Slutprodukt | 12 |
| 3 Termextraktion | 13 |
| 3.1 Manuell termextraktion | 13 |
| 3.2 Automatisk termextraktion | 13 |
| 3.3 Tidigare studier | 15 |
| 4 Jämförelse av manuell och automatisk termextraktion | 16 |
| 4.1 Data | 16 |
| 4.2 Den manuella metoden | 17 |
| 4.3 Den automatiska metoden | 18 |
| 5 Resultat | 21 |
| 5.1 Resultat för den manuella utvärderingen | 21 |
| 5.2 Resultat för den automatiska utvärderingen | 22 |
| 6 Diskussion | 24 |
| 6.1 Diskussion om manuell utvärdering | 24 |
| 6.2 Diskussion om automatisk utvärdering | 25 |
| 7 Avslutning | 26 |
| 7.1 Vidareutveckling | 27 |
| Litteraturförteckning | 28 |

Tack

Först och främst skulle jag vilja tacka mina handledare, Beata Megyesi från Uppsala Universitetet och Henrik Nilsson från TNC. Ett stort tack för all hjälp som jag har fått för denna utvärdering. Alla tips och idéer har varit till stor nytta och även alla genomläsningar och förbättringsförslag. Det har varit givande att träffa er gång på gång och diskutera mitt arbete, både genom personliga möten och genom e-brevväxling. Våra möten har varit nyttiga för mig och dessa har lett mig på rätt väg. Jag skulle även vilja tacka Päivi Pasanen för alla förklaringar och för en liknande undersökning som har inspirerat mig en hel del och bidragit med mycket värdefull information igenom hela uppsatsen. Vad gäller teknisk assistans så ska Trados Support och i synnerhet Anne Katrin Welp från supportavdelningen ha ett varmt tack för alla timmar som spenderades för att få verktyget att producera i minsta detalj det som vi var ute efter. Det krävdes lång tid för att få allting att fungera såsom vi ville, vilket vi till sist lyckades med också. Henrik Nyh förtjänar också eloge för sina kommentarer, sitt stora engagemang i mitt arbete och alla goda råd som jag har fått av honom på vägen. Joel Sundblad likaså, för stort moraliskt stöd, för korrekturläsning och för hjälpen att krydda till utvärderingen med det där lilla extra som gör det hela mycket roligare att läsa. Slutligen vill jag tacka min familj för att ha motiverat mig genom hela arbetsprocessen. Tack ska ni ha.

1 Introduktion

Språk- och datavetenskap förenas inom det tvärvetenskapliga ämnet språkteknologi där syftet är att förenkla och förbättra informationsöverföringen som sker mellan människor och datorer. Språkteknologi är ett brett område med en mängd delområden och resurser som används inom arbetslivet. Ett viktigt sådant delområde är lexikografi där den centrala resursen är elektroniska lexikon. Anledningen till att elektroniska lexikon ses som angelägna är många och en av dem är för att språket är produktivt, det vill säga att nya ord kommer in i språket ständigt - vissa lånade, andra nyskapade.

Med hjälp av elektroniska lexikon kan man slå upp ord, kontrollera stavning och böjning, snabbt få fram ord inom ett visst ämnesområde (beroende på hur mycket lexikonet täcker), lyssna på orden, med mera. Elektroniska lexikon framställs bland annat inom terminologin. Terminologin tjänar till att underlätta kommunikationen mellan personer som är insatta i ett specifikt område. Behovet för utveckling av automatiska metoder märks klart när det gäller extraktion, det vill säga, processen att dra ut terminologisk information från texter. En terminolog, som analyserar, strukturerar och sedan beskriver kunskapen om olika fackområden genom att först och främst analysera och strukturera begrepp, men även en översättare, vill skapa och kompilera sina ordlistor då ständigt nya termer uppstår. Detta kan man göra med hjälp av en språkteknologisk tillämpning, nämligen termextraktion.

Termextraktion är ett delmoment inom terminologiarbetet då man går igenom texter och försöker hitta och samla så många termer som möjligt kring ett visst begrepp. Med denna information ska man därefter studera vilka slags kontexter termerna kan förekomma i och slutligen definiera begreppet. Med hjälp av dagens moderna informationsteknik kan man genomföra vissa delar av termextraktion även på ett automatiskt sätt.

Sedan 1980-talet har ingenjörer och lingvister skapat en hel del automatiska termextraktionsverktyg. I och med att verktygen har kommit så pass sent är detta ett nytt, ungt och ännu outvecklat område. Dessa automatiska termextraktionsverktyg utvärderas ofta genom att man mäter deras precision och deras täckning av termkandidatlistor som verktyget i slutändan producerar.

1.1 SYFTE

Arbetet syftar till att jämföra en termextraktion som är gjord manuellt med en termextraktion som är gjord automatiskt och se hur de olika metoderna fungerar när det gäller att hitta termer i text. Undersökningen ska även visa hur den automatiska termextraktionen kan förbättras i framtiden. Vi vill få fram hur mycket dessa två metoder skiljer sig åt. För- och nackdelar analyseras samt förslag på problemlösningar och utvecklingsmöjligheter ges.

Syftet med detta examensarbete är att göra en utvärdering av Trados termextraktionsverktyg MultiTerm Extract på ett material som är på svenska, innehållandes medicinska termer, där materialet även har extraherats manuellt, tillsammans med Terminologicentrums (TNC:s) terminologer. TNC är Sveriges nationella centrum för terminologi och fackspråk där man bland annat arbetar med att reda ut oklara begrepp. Trados är ett företag som bland annat utvecklar terminologi- och översättningsverktyg.

Det är inte förvånande att en dator utför jobbet snabbare än en människa, men i denna jämförelse kommer kvalitén av utdatan från respektive metod att ha en avgörande roll.

Materialet som den manuella och automatiska excerperingen baseras på är texter ur ”Läkartidningen”, vilket är praktiskt av flera skäl: den är en facktidskrift som normalt innehåller termtäta texter från ett specifikt område, och materialet finns tillgängligt i ett digitalt arkiv.

1.2 STRUKTUR

I kapitel 2 presenteras terminologins teori och hur terminologiarbete går till. Avsnittet inleds med terminologiläran och fortsätter därefter med hur terminologer, med hjälp av experter inom det specifika fackområdet där terminologin ska användas, arbetar steg för steg.

I kapitel 3 ges allmän information om manuell och automatisk termextraktion. I detta kapitel beskrivs de automatiska system som är aktuella i dagsläget, i synnerhet det system som Trados MultiTerm Extract använder sig av.

I kapitel 4 framställs och genomförs den manuella och den automatiska termextraktionen som också jämförs med varandra.

I kapitel 5 presenteras resultatet av den manuella och den automatiska termextraktionsmetoden. Detta kommer att redovisas med en jämförelse.

I kapitel 6 analyseras resultatet. Tidigare studier av andra språk presenteras, diskuteras och resultaten jämförs med varandra. För- och nackdelarna med de respektive metoderna beskrivs också.

I kapitel 7 ges en sammanfattning, några eftertankar, samt förbättringsidéer till hur man kan bygga vidare på de resultat som verktyget ger oss i dag och göra det värt att använda sig av i framtiden.

2 Terminologi och terminologiarbete

Språkets allra viktigaste funktion är att det ska fungera som kommunikationsmedel i vardagslivet. För att kommunikationen inom ett visst fackområde ska vara så bra som möjligt krävs det att det existerar ett utvecklat och korrekt fackspråk som fungerar och utgör en enhetlig beskrivning, där tvetydigheter helst inte får förekomma alls. Därmed ställs högre krav på den fackspråkliga kommunikationen, jämfört med den allmänna. Något perfekt fackspråk finns inte i verkligheten i dag (TNC och Spri, 1999). Ett och samma ord kan betyda helt olika saker för olika personer, även om de arbetar inom ett och samma fackområde.

Ytterligare problem förekommer då olika termer betyder samma sak, även här för människor inom ett och samma fackområde. För att kunna enas om vad ett fackord ska betyda är man tvungen att ena sig om och bestämma ordens betydelse och se till att just denna betydelse förblir konstant och inte ändras i något avseende eller vid något kommunikationstillfälle. Här kommer terminologi in då facktermer behövs för att man ska kunna uttrycka sig så precist och noggrant som det bara går inom sitt ämne och se till att kommunikationen flyter på utan hinder med andra inom samma område. Nedan visas 2 definitioner av terminologi.

”Terminologi är en mängd ord, uttryck och termer som är specifika för ett visst fackområde. Terminologin tjänar till att underlätta kommunikationen mellan personer som är insatta i området. Samtidigt kan det försvåra för nykomlingar. Terminologier uppstår vanligen spontant när behovet finns.”

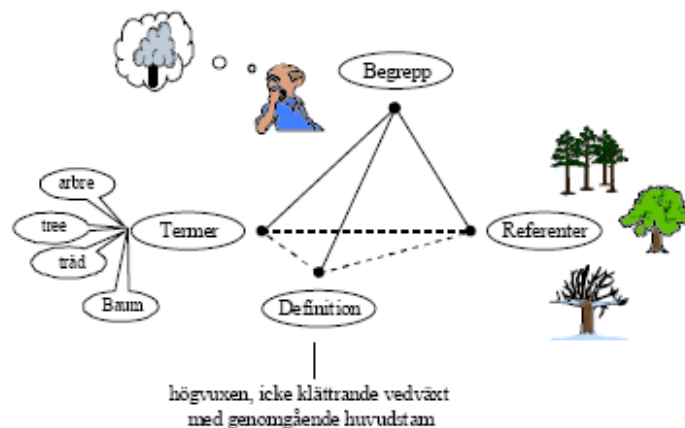
– (Wikipedia terminolog, 2007)

”Terminologi: uppsättning benämningar som hör till ett fackspråk”

– (Nordterm, 2005)

2.1 TEORIN

Inom terminologiläran vill man beskriva, strukturera och överföra kunskap från flera olika teorier. Den är ganska ny då den har utvecklats från 1920-talet i takt med att industrialiseringen, specialiseringen och internationaliseringen har ökat de senaste decennierna. Som lärans fader måste man nämna teknologie doktorn Eugen Wüster som i sin avhandling ”Internationale Sprachnormung in der Technik (1931)” drar upp riktlinjerna för en allmän terminologilära. Hans teori skulle täcka alla fackområden och alla fackspråk. När det gäller den terminologiska analysen var han den förste som satte själva begreppet i centrum.



Figur 1. Tetraeder (TNC, 2007)

Figur 1 (TNC, 2007) visar beståndsdelarna av terminologiarbetet och deras relationer till varandra. Modellens fyra dimensioner tillhör tre olika världar:

- a) Verklighetens värld – handlar om referenter och deras egenskaper
- b) Tankevärlden – handlar om begrepp och deras särdrag som återspeglar referenterna
- c) Språkvärlden – handlar om termer som namnger begrepp samt definition som beskriver begrepp och som skiljer dem från varandra.

Referenter kan exempelvis vara föremål, händelser, egenskaper med mera, där vissa är konkreta och andra abstrakta. För att vi ska kunna resonera om denna utomspråkliga verklighet så krävs det att vi uppfattar dess olika beståndsdelar och vidare kategoriserar det hela. Kategoriseringen i sin tur bygger på gemensamma drag hos en mängd referenter.

Exempel 1.1

TRÄD

+ *vedväxt*

+ *har genomgående huvudstam*

Med kategoriseringen kan vi forma ett begrepp, alltså en enkel föreställning om en grupp av referenter. Därmed blir begreppet ”trä” en mental, icke-språklig föreställning om alla möjliga olika träd (TNC och Spri, 1999). Det blir helt enkelt en avbildning som vi föreställer oss i våra tankar. Begreppets beståndsdelar brukar man kalla för kännetecken och dessa är vanliga tankeskapelser av egenskaperna hos en referent. I vårt fall blir begreppet ”trä” en kombination av kännetecknen: vedväxt, högvuxen, icke-klättrande, med genomgående huvudstam (Suonuuti, 2004).

Begrepp är således endast abstraktioner innehållande materiella och icke-materiella föremål (Vintar, 2005). Då begreppen bara existerar i tankevärlden behöver vi beskriva dem och benämna dem för att vi ska kunna tala och skriva om dem. Vi måste ge begreppen ett liv. Därför är vi tvungna att ge dem termer när vi kommunicerar. Termer är mestadels fackspråkliga ord som används inom ett specifikt fackområde. Termer är även uttryck för begrepp som har en definition. Med vårt exempel 1.1 blir begreppet ”trä” försett med termer på olika språk. I och med termen kan vi förflytta oss ytterligare ett steg från själva verkligheten, det vill säga, från referenterna, sedan från en kognitiv nivå, alltså, från begreppet, och in i det språkliga. Allt som oftast är vi många som ska vara överens om hur en term ska användas. Det som är allra viktigast då är att göra en bra definition av det begrepp som termen står för.

Definitionen av ett begrepp brukar lösa många framtida och långsiktiga problem, där är den en jättevägvisare bit av pusslet. Med definitionen ger man en språkligt konstruerad beskrivning av ett begrepp som bygger på begreppets kännetecken och som ska skilja begreppet mot de övriga. De heldragna linjerna i figur 1 visar att alla relationer mellan term–referent, term–definition och definition–referent sker genom begreppet. Därmed vågar jag mig på slutsatsen att begreppet finns konstant närvarande även om ett eller flera andra hörn inte jämnt är med. Som exempel hittas ingen etablerad term för de små pappersbitar som kommer från hålslaget och ingen referent bakom begrepp som ”enhörning” och ”tomten”. Vidare finns det många begrepp som saknar en riktig definition inom etablerade fackområden. Ett sådant fackområde är politikernas språk, där jag fann att ordet *region* används olika i olika kontexter.

Exempel 1.2 (Wikipedia region, 2007)

REGION

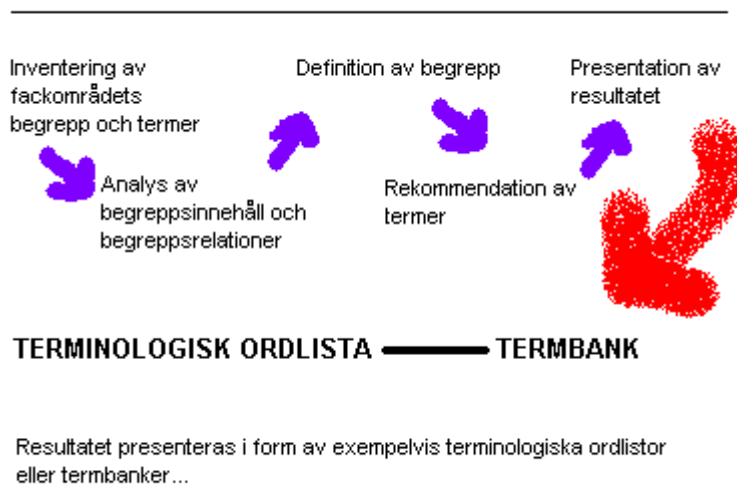
+ *landstinget i Skåne län*

+ *geografiskt område*

Slutsatsen blir att begreppet är själva kärnan för allt terminologiarbete.

2.2 TERMINOLOGIARBETET

I de flesta fall definieras ordet *terminologi* som en samling av ord och uttryck med särskild betydelse på ett specifikt område. Men en *terminologi* kan även tolkas som läran om sådana terminologier som är ett resultat av ett terminologiarbete. Mitt manuella upplägg för terminologiarbetet av denna utvärdering visas i figur 2:



Figur 2. Terminologiarbetet från början till slut

Därmed har terminologiarbetet ett praktiskt syfte där det hela går ut på att åstadkomma en effektiv fackspråklig kommunikation och att minimera så mycket det bara går av alla allvarliga och kostsamma missförstånd men även av tidsödande diskussioner.

Det handlar om att på ett systematiskt sätt, med hjälp av metoder som är sammanställda inom terminologiläran, samla, analysera, beskriva och presentera diverse fackområdets begrepp och deras termer. För att nå ett lyckat resultat arbetar terminologer tillsammans med fackmän. De identifierar och analyserar centrala begrepp inom ett visst fackområde. Sedan rangordnas begreppen och beskrivs med hjälp av definitioner som indirekt anger relationerna till snarlika begrepp inom detta område. I sista steget väljs en, men ibland även flera termer som ses vara lämpliga för varje begrepp ut. Slutligen blir resultatet en terminologi och visas antingen i form av en ordlista eller en termdatabas. Denna process har använts i denna utvärdering, men undantaget att inga fackmän deltog i undersökningen.

Det första man behöver veta för att utföra ett terminologiarbete är vad terminologin ska användas till och till vilken målgrupp. I och med detta är det första steget att avgränsa det terminologiska fältet, vilket innebär att man gör en detaljerad beskrivning av det specificerade fackområdet och att man delar upp det i flera mindre delområden. I andra steget sker förbearbetning av dokument och undersökning av vilka källor som verkligen är pålitliga. Själva avsikten med detta är att välja ut termer ur texter med omfattande kontext som man därefter har som underlag för projektet. Exempel på begreppen som man plockar ut (TNC och Spri, 1999):

- a) fackspecifika begrepp – begreppen är specifika då de används inom det fackområde som behandlas i fråga (t.ex. ”telekommunikation” inom teleområdet).
- b) facköverskridande begrepp – begrepp som kan placeras i två eller flera områden (ex. ”modul” som hör hemma inom byggområdet men även inom dataområdet med flera).

- c) lånade begrepp – begrepp som konstant används inom fackområdet i fråga, men hör ändå för det mesta hemma inom ett helt annat område (t.ex. ”dataskärm” & ”acklimatisering” inom arbetarskyddsområdet).
- d) allmänna begrepp – begreppet som används allmänt och som inte direkt kan klassificeras som hörande till ett specifikt fackområde (ex ”system” & ”process”).

Det tredje steget som en terminolog mer eller mindre måste göra tillsammans med en expert inom det specifika fackområdet är att skriva definitioner till begreppen. Ett visst antal regler gäller under detta definitionsarbete. Definitionens innehåll är viktigt eftersom en definition måste beskriva begreppet och inte en viss referent eller liknande. Håller vi oss till vårt exempel 1.1 med begreppet ”träd” innebär det att om en definition beskriver detta begrepp så beskriver den endast den mentala föreställningen som vi själva har bildat då vi vet alla de gemensamma dragen hos alla träd och inte bara de som finns hos exempelvis en tall.

Definitionen ska inledas med ett överbegrepp till begreppet som ska definieras. I vårt fall med begreppet ”tall” väljer man överbegreppet ”barrträd” som är närmast överordnad:

- (1) TRÄD
- (2) BARRTRÄD
- (3) TALL

Vidare krävs det att definitionen inte är för lång och inte heller för komplicerad. Dessutom ska den följa alla de övriga reglerna som vi har nämnt tidigare när det gäller allt definitionsarbete. Viktigast av allt måste definitionen anpassas till den specifika gruppen och inte ha med begrepp som dessa användare inte förstår. Av detta kan vi dra slutsatsen att av de alla terminologiska arbetsmetoderna är begreppsanalysen den viktigaste.

Det fjärde steget handlar om att välja bra termer för definitionen av begreppet. En terminolog i dag använder sig mer eller mindre av följande sju krav för att skapa en bra term.

I. Termen ska vara precis. Exempelvis, en spak som öppnar en lucka ska hellre heta öppningsspak än öppnare.

II. Termen ska vara entydig för varje ämnesområde. Detta innebär att termen inte får vara en polysem och inte heller en homonym. En polysem är ett ord som har flera relaterade betydelser och en homonym är ett ord som både stavas och/eller uttalas likadant men som har olika betydelse. Polysem och homonym illustreras i exempel 1.10.

Exempel 1.10

Polysem term:

salta

1. konservera livsmedel med tillsats av salt
 2. smaksätta genom att strö lite salt över livsmedel eller färdig maträtt
- (ur: Matlagningstermer)

Homonym term: **1 fil**
enkelriktad, vanligen markerad, del av bred körbana
2 fil
filmjolk
(ur: Svensk ordbok)

III. Termen ska vara accepterad av fackmän, i vissa fall även internationellt. Finns det redan en etablerad term så behöver man inte hitta på en ny bara för sakens skull. Ett exempel är ”pacemaker” som gjorde en sådan revolution att den blev fångad i det svenska språket omedelbart och man gjorde ett tappert försök med ”hjärtstimulator”, men den svenska termen slog inte originalet i alla fall.

IV. Termen ska passa in i användarens språksystem vad gäller stavning, böjning och uttal. Som exempel kan man nämna att den svenska termen ”gränssnitt” passar mycket bättre i det svenska språksystemet, enligt de kriterierna som används, än om vi hade lånat in engelskans ”interface”.

V. Termen ska inte vara missvisande. Ett typiskt exempel på detta är 60-talet då både ”atomenergi” och ”kärnenergi” förekom. Man tacklade problemet genom att visa att förledet ATOM- är något som rör hela atomen och förledet KÄRN- är något som rör atomens kärna, som i sambandet med klyvning av en kärna. Därmed blev svaret ”kärnenergi”, vilket faktiskt används mest frekvent i dag.

VI. Termen ska vara språkekonomisk, vilket innebär att man hellre vill ha en kortare term än en längre. Exempel på det är ”Betalkort” och ”Betalningskort” – här anses ”betalkort” vara ett bättre val.

VII. Termen ska vara genomsynlig och i och med detta ska man kunna förstå enkelt vad den betyder. Även om man exempelvis inte jobbar på posten så är så kallade posttermerna ”höghusbrevbäring” och ”dörr-till-dörr-service” lätta att förstå.

En terminolog försöker alltid att följa dessa sju ovannämnda krav, men jag tycker inte att det alltid går. Ibland motsäger den ena den andra som i exemplet ”jitter” vs. ”höghusbrevbäring” där ”jitter” är språkekonomisk men inte genomsynlig och ”höghusbrevbäring” är precis tvärtom. Riktlinjerna säger att om det finns flera termer för samma begrepp så ska man välja en av dem och ha den utvalda som den rekommenderade termen. Man anger då den rekommenderade termen som uppslagsord och de övriga termerna som synonymer. Om valet står mellan en term och en förkortning så är den oförkortade formen bäst att använda, med undantag när en förkortning är mer känd i det specifika fackområdet än den oförkortade formen (ex. *plocken* i stället för *plockmaskinen*, när det gäller lagerarbete), för då föredrar man förkortningen som den rekommenderade termen.

Det femte och sista steget som man gör i ett terminologiprojekt är att presentera arbetet. Detta gör man vanligast med terminologiska ordlistor eller termbanker. Där lägger man fram den terminologiska informationen i form av termposter och varje sådan innehåller information om bara ett enda begrepp. Informationen som man inkluderar brukar variera, men det som oftast brukar vara med är rekommenderad term, användningsområdet, synonymer, definition, eventuell anmärkning, hänvisning till relaterade begrepp, administrativa uppgifter ifall det behövs och ekvivalenter på andra språk.

2.3) SLUTPRODUKT

När man väl är klar med sitt tillvägagångssätt, vare sig man har valt det manuella eller det automatiska, är det hårt arbete som väntar med att bestämma varje begrepps riktiga betydelse och innehåll.

Arbetsprocessen är lång där man har många frågor att besvara. Det hela går ut på att hitta ett begrepps avgörande kännetecken som avgränsar dess omfång och innehåll från de övriga begrepp. Det kommande arbetet illustrerar jag i exempel 1.13 med ordet besök.

Exempel 1.13

BESÖK

- (a) Jag kan vara på besök hos mina nära och kära då jag hälsar på dem i deras hem.
- (b) Som kurator på Westerlundiska Gymnasiet i Enköping, exempelvis, kan jag ta emot besök, men det är inte alls samma slags besök då innebörden är insnävat i förhållande till (a).

När man, i vårt fall med personal på en skola, ska försöka ge begreppet ”besök” en så korrekt och avgränsad mening som ska gälla inom området, måste man ta fram alla kännetecken som begreppet har som fackterm inom det specifika området. Slutligen skulle jag vilja skriva definitionen:

”Personlig kontakt inom utbildning som innebär personligt möte mellan elev och rådgivare”.

3 Termextraktion

Termextraktion är en betydelsefull del av terminologiarbetet. Det som görs inom termextraktion är att man i allmänhet går igenom en större textsamling eller en vanlig text och försöker hitta begrepp och deras benämningar, såsom termer och förkortningar. Man noterar all relevant information om begreppet i fråga (The Pavel Terminology Tutorial, 2006). Informationen handlar oftast om hur begreppet definieras och om dess kontexter. Givetvis också hur begreppet används i alla möjliga kommunikationer i vardagslivet, även inom det skriftliga språket. Inom terminologin kallas de samlade texterna som man väljer ut (där alltså begreppen eller de språkliga uttrycken förekommer) för excerpter.

Excerpterna kan hämtas ur alla möjliga slags texter. Oftast tar man dem ur antingen terminologiska ordböcker eller ur lexikografiska ordböcker. Andra exempel på material är handböcker och lagtext men även webbinformation, fast där får man vara extra försiktig när det gäller att lita på källor och innehåll.

Termextraktion kan göras på olika sätt, manuellt och automatiskt. Nedan beskrivs de utmärkande dragen för respektive metod.

3.1 MANUELL TERMEXTRAKTION

Till en början görs en manuell termextraktion av en människa alltid för hand. Det första som sker är identifiering av termer, som är den delen av termextraktion som har att göra med igenkänning och val av termer. Detta innebär att man går igenom textmassan och väljer ut termer som sedan ska studeras ytterligare och förhoppningsvis spridas ut i verkligheten som riktiga användbara ord, förvisso till de specifika områden där dessa ord behövs. Sedan skrivs definitioner med hjälp av relevant information om begreppet i fråga. När definitionen är klar krävs det expertgranskning för att försäkra att dessa termer faktiskt är en del av det specifika språket och att begreppen verkligen tillhör fackområdet. Först efter begreppsanalysen börjar överväganden om de existerande termkandidaterna och eventuellt tas några bort – i bästa fall, alla utom en, som blir den rekommenderade termen för begreppet i fråga. Den manuella termextraktionen brukar allt som oftast göras i samråd med olika experter från de olika specifika områdena. Så går man till väga för att resultatet ska bli så bra som möjligt, av den enkla anledningen att i slutändan är det just de experterna, och andra människor som arbetar inom deras område, som ändå ska använda termerna.

3.2 AUTOMATISK TERMEXTRAKTION

Automatisk termextraktion ur en inläst textsamling är ett utmärkt och ett snabbt sätt att skaffa sig kunskap om ämnesområdet samt det specifika språket och skulle kunna utgöra ett viktigt sätt att effektivisera denna del av terminologiarbetet.

Man kan antingen göra en enspråkig eller flerspråkig termextraktion. Vid en enspråkig försöker man analysera en text för att identifiera termkandidater, medan man vid en flerspråkig analyserar existerande källtexter med deras översättningar för att känna igen termkandidater och deras ekvivalenter. Trots att den inledande extraktionen genomförs av ett datorprogram, måste den resulterande matchningslistan bekräftas av en mänsklig terminolog eller översättare. I och med detta är den automatiska termextraktionen mer begränsad än den manuella termextraktionen. Därför bör termextraktion betraktas som en datorstödd snarare än som en helautomatisk process (Bowker 2003, Esselink, 2003).

Automatisk termextraktion kan basera sig på en lingvistisk eller en statistisk metod, men i vissa enstaka fall av en kombination av dem båda. Trados MultiTerm Extract gör ett försök att kombinera dessa metoder (SDL MultiTerm Extract, 2007).

En lingvistisk metod strävar efter att identifiera ordkombinationer som matchar vissa specifika ordklass-mönster (t.ex. "adjektiv + substantiv" eller "substantiv + substantiv"). Metoden är språkspecifik och kräver ofta omfattande språkliga resurser. Den typiska lingvistiska metoden för termextraktion är att plocka ut ordföljder som motsvarar vissa termtypiska ordklassmönster. I svenskan är, som nämndes i exemplet ovan adjektiv – substantiv ett vanligt mönster för termer medan preposition – artikel inte alls förekommer lika ofta, om någonsin. Bland annat Justeson och Katz (1995) beskriver en lingvistisk filtrering där källtexten ordklasstaggas, alltså, där allt som inte motsvarar en giltig uppsättning av taggmönster filtreras ut. Det krävs även att kandidattermen har en viss minimifrekvens. För nominalfraser gäller det reguljära uttrycket:

$$(((A|N) + |((A|N) * (NP)?)(A|N) *))? N$$

N = substantiv

A = adjektiv

P = preposition.

? = det som kommer precis före är frivilligt. (NP)? = ((NP)) = NP eller ingenting

Det reguljära uttrycket accepterar exempelvis ”giltig term” (ADJEKTIV-SUBSTANTIV) medan ”av den” (PREPOSITION-ARTIKEL) tas direkt bort.

När det gäller resultatdelen, precisionen och täckningen, så brukar taggmönstren ge god precision, fast delvis på bekostnad av täckningen. Över 80 % av termer är nominalfraser (Arppe, 1995). Därmed kan det tyckas rättvist att endast fokusera på ordklassmönster som motsvarar nominalfraser, vilket man också ha gjort.

En statistisk metod söker efter upprepade sekvenser av lexikala element. Statistiska metoder för termextraktion tar fasta på de statistiska egenskaper som skiljer termer från ord i löpande text. Det vanligaste är att fokusera på flerordstermer och titta på associationsmått för de ingående orden (Algeria, 2004). Sådana associationsmått bygger ofta på det informationsteoretiska måttet som kallas för *ömsesidig information* och som definieras (Church och Hanks, 1989):

$$I(x, y) = \log^2 (P(x, y) / P(x)P(y))$$

Det man jämför är sannolikheten att två händelser inträffar tillsammans. Ordförekomster blir händelser i vårt fall med termextraktion. Huvudpoängen är att man jämför sannolikheten för att de inträffar oberoende av varandra. Ifall båda sannolikheterna är lika stora blir kvoten cirka 1 och logaritmen cirka 0. En större förenad sannolikhet innebär ett högre informationsvärde. Värdet kan bli negativt om orden aldrig förekommer tillsammans.

Frekvenströskeln, som indikerar antalet gånger ett ord eller en ordsekvens måste upprepas för att kunna betraktas som ett termalternativ, kan ofta specificeras av användaren. Den största fördelen med den statistiska metoden är att den är språkoberoende till skillnad från den lingvistiska metoden. Å andra sidan ger den lingvistiska metoden bättre avgränsade termer och färre repeterade böjningsformer än den statistiska metoden (Bowker 2003, Esselink, 2003).

Exempel på automatiska ingående faser som används vid lingvistiska och statistiska metod:

- a) Tokenisering – man delar upp texten i enheter som kallas ”tokens” där ett ”token” är en sträng av bokstäver och tecken som skiljs åt av mellanslag.
- b) Stoppordslista – man använder sig av en lista som består av ord som är så vanliga att de saknar särskiljande förmåga; dessa ord är alltså inte termer. Denna ingående fas använder bland annat termextraktionsverktyget Trados MultiTerm Extract med en lista på 306 stoppord i vilka bland annat ingår substantiv, verb, pronomen och prepositioner.
- c) Stemming – Man tar fram stammen på varje ord. Stemming används för att öka täckningen på bekostnad av precisionen. Exempelvis, om vi gör en sökning på ”blommor”, stemmas det till ”blomm”. Om det vi söker i har stemmats också, kan vår sökning hitta ”blommans”, ”blommorna”, etc. också.

Bland annat SDLPhraseFinder (2005) använder sig av den lingvistiska metoden och System Quirk (2007) är ett exempel på ett verktyg som använder sig av den statistiska metoden.

3.3 TIDIGARE STUDIER

En studie liknande den här utvärderingen har gjorts med det finska språket av Päivi Pasanen (2005). Hennes syfte var att ta reda på hur hjälpsamt ett termextraktionsverktyg är när det gäller de finska termerna. Till en början bad Pasanen en sjöfartsexpert att excerpera termer ur en finsk källtext som innehåller 2 366 ord (Wihuri, 2002). Sedan användes NaviTerm 2.0 (2007) och Trados MultiTerm Extract som automatiska verktyg. Grundinställningar användes i båda verktygen med avvikelse att ”noise ratio” ställdes på 75 % i MultiTerm Extract. Ju högre procentgrad noise ratio ställs på desto fler termkandidater hittas. Pasanen gjorde ett försök att få ut fler finska termkandidater än vad man får med 50 % ”noise ratio” som grundinställning.

Resultatet blev att 220 termkandidater togs ut för hand medan verktyget MultiTerm Extract hittade 306 stycken. Av de 220 manuellt plockade termkandidaterna var nästan 70 % av dem termkandidater som förekom endast en gång i texten. MultiTerm Extract hade flera problem med det finska språket, speciellt med just de termkandidater som förekom endast en gång i texten. Där hittade verktyget upp mot 90 % sådana termkandidater, det vill säga 268 av det sammanlagda antalet 306.

I jämförelsen hittade verktyget 50 termkandidater av de 220 som experten hade tagit ut. Därmed är täckningen på 23 %. Av 306 termkandidater var endast 50 stycken korrekta, vilket ger en precision på 16 %. Sjöfartsexpertens termkandidatlista innehöll 68 (av 220) stycken termer som förekom två gånger eller fler i texten. I MultiTerm Extract hittades det 184 stycken förekomster. Av dem var det 39 stycken som var giltiga och detta ger oss en precision på 21 % och en täckning på 57 %. Detta innebär att verktyget excerperade en hög frekvens av ord och fraser som i själva verket inte är termer.

Pasanen var inte nöjd med MultiTerm Extract, när det kom till det finska språket. Förvisso hittade verktyget runt 50 % av de två ords termkandidater som förekommer i texten. Men det är ändå 52 av 64 tvåordstermer som aldrig excerperades. Dock är verktyget förträffligt när det gäller att hitta termer som har frekvensen 3 eller högre. När man studerar resultatet så inser man klart och tydligt att MultiTerm Extract både övergenererar (endast 50 av 306 termkandidater var giltiga) och undergenererar (endast 12 av 152 termer som förekommer en gång i texten hittades) i det finska språket. Pasanen tror dock att verktyget ska fungera bättre för bland annat de germanska språken.

4) Jämförelse av manuell och automatisk termextraktion

Metoderna som ska användas för en så utförlig utvärdering som möjligt är den manuella och den automatiska. Den manuella görs i samarbete med terminologer från TNC och den automatiska med hjälp av programmet Trados MultiTerm Extract. Båda metoderna använder sig av samma data.

Ett termextraktionsverktyg utvärderas med avseende på precision och täckning. Då man mäter precision mäter man hur många av de extraherade termkandidaterna som är giltiga termer. Är precisionen 77 % så innebär det att 77 % av termkandidaterna är giltiga termer, vilket leder till att 23 % är ogiltiga. Med täckning mäter man hur många termer som har upptäckts av alla möjliga termer från texten. Är täckningen 77 % så innebär det att 77 % av termerna har hittats i texten, vilket leder till att 23 % saknas.

Generellt brukar precisionen för ett termextraktionsverktyg variera mellan 30 till 70 % (Pasanen, 2005), vilket innebär att i snitt är bara hälften av de excerperade termkandidaterna giltiga termer. Ändå finns det vissa som tror att precisionen kan uppnå 90 % (Soininen, 1999). Anledningarna är många till att dessa verktyg inte har lyckats uppnå högre procentsäkerhet än det som nämns här ovan. Det är nog svårt att veta om en term är en term även för en människa. Men sedan uppstår det även problem med över- och undergenerering. Inom terminologin innebär övergenerering extraktion av för många termkandidater, där många av dem i själva verket inte är termer. Undergenerering innebär att man har med för få termkandidater och därmed saknas en hel del termer. Hittills har de flesta verktyg resulterat i antingen övergenerering eller undergenerering av termer.

4.1) DATA

Då det svenska språket skulle undersökas används 15 stycken artiklar från Läkartidningen som data för projektet. Totalt innehåller de 14 147 ord. Läkartidningen fungerar utmärkt som data för den här uppgiften då den innehåller många återkommande termer som används inom ett specifikt område. Dessutom finns hela materialet som används i detta projekt ute på nätet så det går under alla omständigheter, när som helst, att kontrollera allting på egen hand. Dessa artiklar finns alltså som digitalt arkiv och man kommer åt dem genom att söka på dem på webbsidan: <http://larkiv.lakartidningen.se>

Artiklarna som användes finns listade i tabellen nedan.

| TITEL | ÅR | TIDNINGSS NUMMER | SIDA |
|--|------|---------------------|-----------|
| Trippelterapi vid dyspepsi hjälper inte 70 procent | 1999 | 6 | 584 |
| Höga nivåer av serumretinol ökar risk för framtida fraktur | 2003 | 12 | 1026 |
| Tamoxifen och magnetfältpåverkan | 2006 | 8 | 15 |
| Litium - ett terapeutiskt grundämne ännu oöverträffat som profylax vid BPD | 2004 | 5 | 376-377 |
| Studier av sömnapné dåliga men andningsproblemet stort | 1997 | 16 | 1502 |
| Expertuttalande om havre i behandlingen av celiaki i Sverige | 1999 | 50 | 5606 |
| Är psykiatrisk funktionskränkning fusk? | 2004 | 51 | 4246 |
| Osäkerhet kring studie av fluoxetin till gravida | 1996 | 42 | 3670 |
| Celiaki hos pappa påverkar det nyfödda barnet | 2001 | 38 | 4060 |
| Ny serie: Serotonin och känslor | 1997 | 38 | 3245 |
| Implementering av intensiv blodglukos- och blodtrycksregim hos diabetiker i England ger marginell merkostnad | 2003 | 7 | 504 |
| Akupunktur kan ge kärlskador | 1998 | 3 | 180-181 |
| Vart tredje barn mår dåligt av sjukhusvistelse med narkos | 2007 | 3 | 120 |
| Aspergers syndrom - vägen till en diagnos | 1996 | 36 | 3788-3794 |
| Valgisering eller valgusering - valet är inte fritt | 2003 | 20 | 1816 |

Figur 3. Läkartidningensartiklar

Det fanns även svårigheter med denna data. Enbart den text som förekommer i själva artikeln extraherades. Texten som står därefter, i de flesta fall på samma blad, räknas inte som data, utan tillhörde andra artiklar.

En nackdel med MultiTerm Extrakt när det gäller att hämta in dessa filer är att verktyget inte klarar av att öppna PDF-filer. Därmed kopierades all text från PDF-filerna och klistrades in i en Word-fil och sparades i doc-format. Manuell rättning utfördes för att se till att allt som ska vara med i utvärderingen kom med i Word-filen, vilket givetvis tog tid att genomföra och kontrollera.

4.2) DEN MANUELLA METODEN

Den manuella metoden innebär att man gör terminologiarbetet för hand och helst tillsammans med någon expert från det specifika området. Någon expert från läkarbranschen närvarade tyvärr inte i detta experiment.

Den manuella termigenkänningen brukar oftast vara väldigt individuell. Det som är en term för en person behöver inte vara en term för en annan. Detta problem minskas radikalt om man genomför uppgiften med hjälp av flera personer, vilket gjordes i denna utvärdering, av cirka 10 annoterare.

Så här löd instruktionerna till alla annoterare, för att göra det så enkelt som möjligt: ”Stryk i den markerade texten under (eller över, med överstrykningspenna) alla ord och uttryck du anser vara termer. Obs! Utgå inte från vad du skulle välja till TNC-bas! Jämför inte heller med någon annan källa. Skriv namn på detta blad – och gärna också några kommentarer om vad som gjorde att du valde dessa termer”.

Denna metod genomfördes enligt de riktlinjer som beskrevs tidigare i avsnitt 2.2, som handlar om just terminologiarbete, eftersom det är på det sättet manuell termextraktion går till.

4.3) DEN AUTOMATISKA METODEN

För utvärderingen av den automatiska metoden valdes termextraktionsverktyget Trados MultiTerm Extract. Detta med anledning av att verktyget kan excerpera termkandidater, deras eventuella översättningar och presentera dem i en termkandidatlista (User Guide MultiTerm 7 Extract, 2005). Man kan excerpera termer ur enspråkiga eller tvåspråkiga dokument, men också från översättningsminnet som är lagrat i systemet. Ur den termkandidatlistan kan man sedan bekräfta att en termkandidat faktiskt är en giltig term. Ett annat val som tillåts är att man kan även stryka termkandidater eller göra ändringar i dem och lägga till ytterligare termer i listan, manuellt, efter att den automatiska termextraktionen har exekverats.

Eftersom det enbart var ett språk som skulle hanteras valdes "Monolingual Term Extraction Project" i menyn.

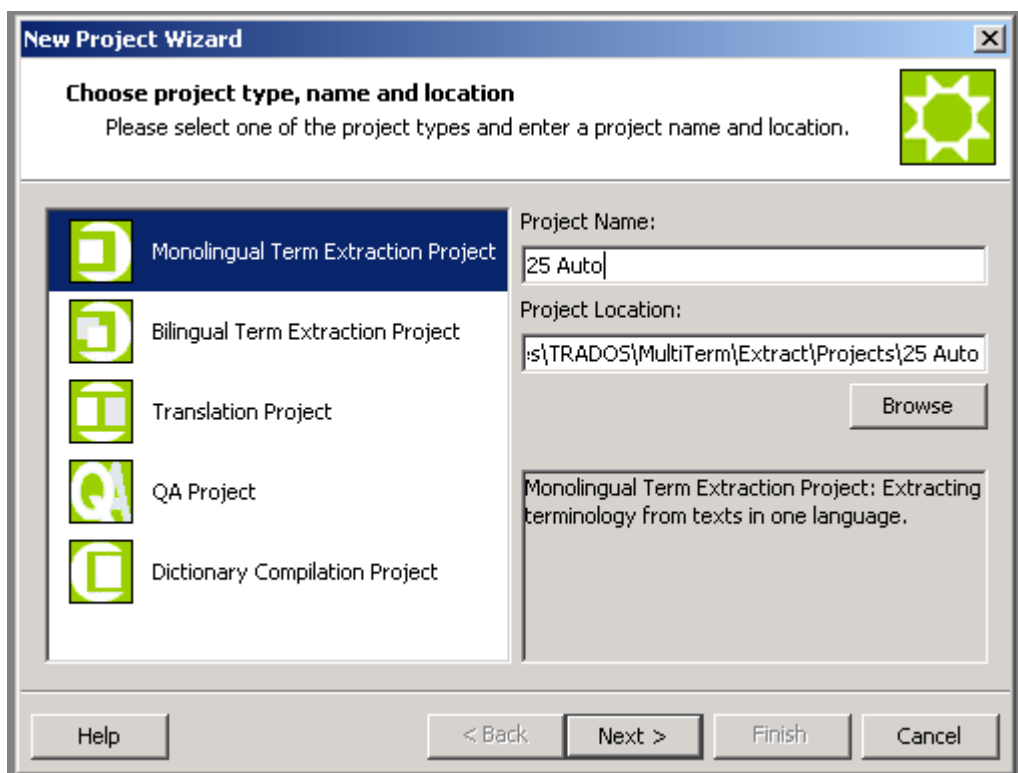


Bild 1. Val av projekt

I nästa steg skulle man välja en termdatabas och språket som ska användas. Det finns i skrivandets stund tyvärr ingen termdatabas för det svenska språket; bara för engelska, tyska, spanska och franska. En termdatabas innehåller inlagda termer i en databas. Texten (datan) laddas in i databasen för att kontrollera om något i texten matchar mot en term från databasen. Därmed tar termdatabasen fram snabbt en lista över funna termer i texten. Då det inte fanns någon termdatabas för det svenska språket kunde endast källspråket väljas.

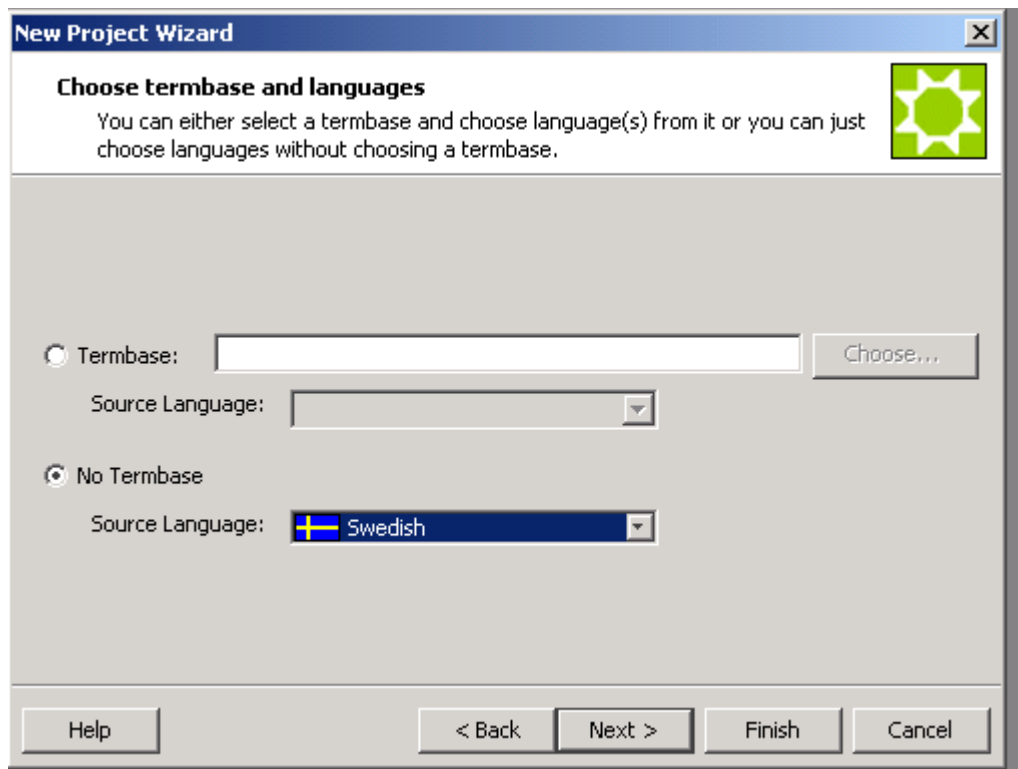


Bild 2. Val av språk

Därefter lägger man upp alla filer som innehåller data som man vill att verktyget ska excerpera. Då programmet inte stödjer PDF-filer som artiklarna var publicerade i på nätet, så klistrades all text in i en Word-fil.

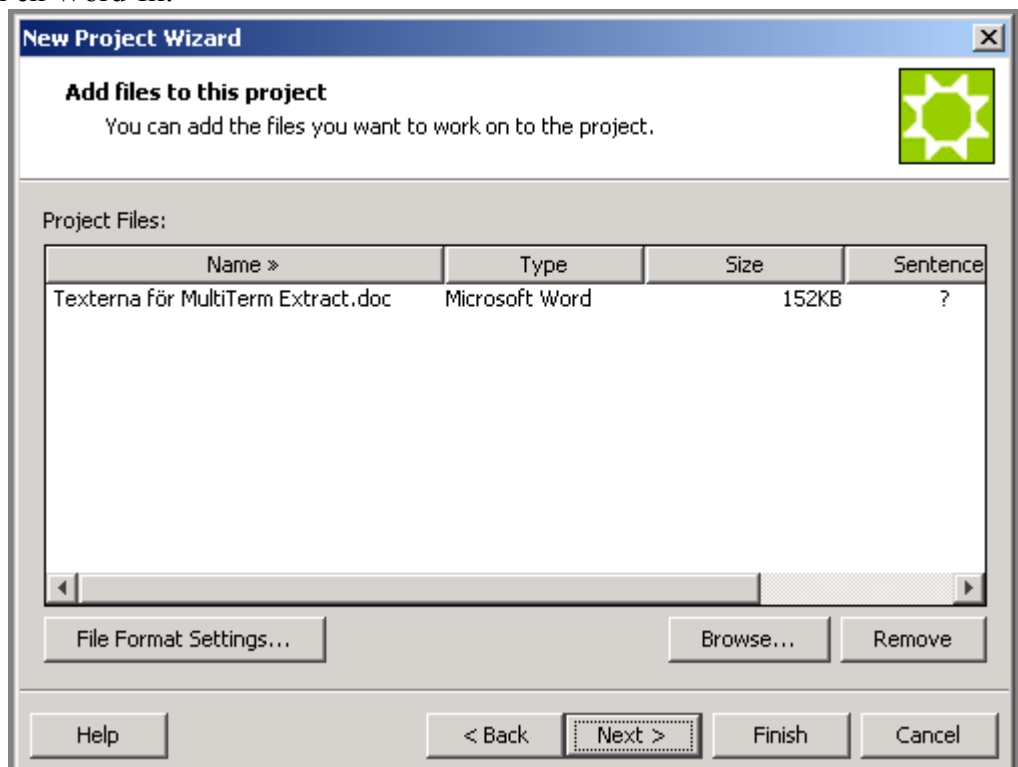


Bild 3. Val av data

Därefter får man en förfråga om man vill att några termdatabaser eller filer som innehåller termer ska ställas utanför. Detta behövs inte i det här projektet, så vi går vidare till nästa steg.

I nästa steg, som även är det sista steget innan verktyget kan sättas igång och utföra själva extraktionen, ska man bestämma det som kallas för "noise/silence ratio" vilket handlar om över- och undergenerering. Ju högre noise ration ställs på desto fler termkandidater hittas och vice versa. Ställs den på exempelvis 10 % så kommer verktyget att ta med enbart de termkandidater som verktyget är 90 % - 100 % säkert på att dessa är termer. Ställs den på exempelvis 90 % så kommer verktyget att ta med alla de termkandidater som verktyget är 10 % - 100 % säkert på att dessa är termer. I det här projektet kommer det att utvärderas tre olika inställningar: 25 %, 50 % och 75 %.

På andra liknande experiment som har gjorts de senaste åren, såsom Pasanens (2005), så har "noise ratio" ställts på högre procentgrad än vad Trados har som grundinställning (50 % på noise ratio och 50 % på silence ratio), oftast på 75 % men även i vissa fall högre. Fördelen med att öka noise ratio är att man får ut betydligt fler termkandidater än vad man får om man använder sig av grundinställningen på 50 %. Nackdelen brukar dock vara att det hela resulterar i att man får mycket "noise" också, det vill säga, termkandidater som i själva verket inte är giltiga termer. I steget efter detta höjdes det maximala antalet termer från 100 till 1 000 stycken, endast säkerhetsmässigt, för att ingenting ska vara begränsat.

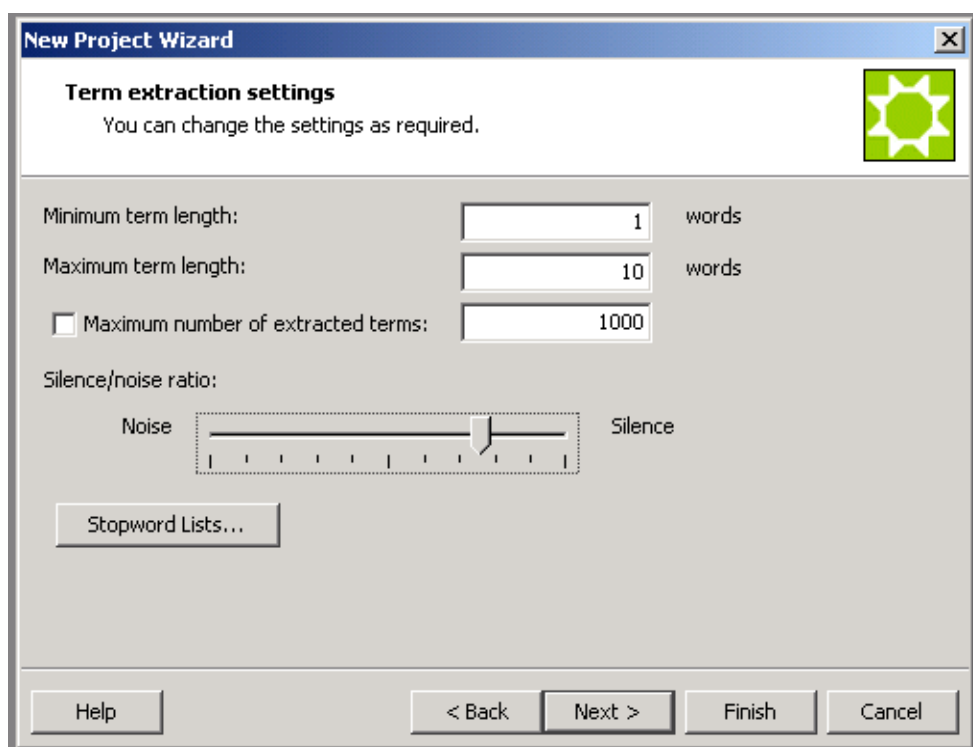


Bild 4. Val av inställningar

Noise ration räknas från väster till höger där varje steglängd är 10 %. På bilden ovan ser vi ett klipp där noise ration är ställd på 75 %. Tidigare nämndes att som ingående termextraktionsfaser som använder statistisk metod använder antingen tokenisering, stemming eller en stoppordslista. I vissa fall kan man även kombinera alla dessa system och därmed köra med flera stycken på en gång, ifall det nu skulle gynna syftet med det specifika projektet. Vi ser också på bilden ovan att Trados har valt att använda sig av en stoppordslista.

Fördelen med att göra termextraktion med hjälp av den automatiska metoden är att det krävs bara en person för hela arbetet och det går snabbt att utföra. Frågan är då hur bra resultatet blir.

5) Resultat

Den manuella utvärderingen gjordes i tre steg och den automatiska utvärderingen gjordes bara i ett steg. Dock ställdes noise ration på 3 olika nivåer så det hela exekverades 3 gånger för att utvärdera resultatet av dem. Här nedan presenteras resultatet av detta.

5.1) RESULTAT FÖR DEN MANUELLA UTVÄRDERINGEN

Allt som allt plockades det ut totalt 654 termer för hand ifrån Läkartidningens texter bestående av 14 147 ord. Mer än 70 % av dessa termer var enordstermer som förekom en gång i texten. I och med att tre granskningar gjordes sammanlagt så lades det till även de termer som av vissa terminologer inte hade tagits med, men ändå tagits med av andra terminologer. Om det var som så att vissa hade extraherat bara ett ord ur en fras och andra fler än ett ord så togs båda fallen med i resultatet. Ifall ett ord ströks under en gång eller flera så räknades det som ett ord i slutändan, i alla fall.

De flesta annoterare på TNC tyckte det var svårt att veta vad som är ”äkta” flerordstermer och vad som är separata termer staplade efter varandra. Ett exempel på detta ur Läkartidningen: ”En patient med arteriosklerotiskt betingad cirkulationsinsufficiens”

Efter en diskussion så kom vi fram att här kan det röra sig om båda fallen, alltså att det är en flerordsterm men att det också kan vara två separata termer, staplade efter varandra. I denna manuella undersökning hade man valt att excerpera adverbet arteriosklerotiskt och substantivet cirkulationsinsufficiens som separata termer, trots allt. Ytterligare ett stort problem som upptäcktes var att den manuella metoden är inkonsekvent på flera sätt och vis. Problemet var att vissa terminologer strök under en term som förekom mer än en gång flera gånger medan andra indikerade den termen bara en gång. Det var även en del andra problem som uppstod, där vissa av dem hade kunnat undvikas om instruktionerna varit tydligare, speciellt det första fallet:

- a) Stamord (t.ex. ischemi) med böjningsformer (t.ex. ischemi-n/r) ska anses vara termer också. I många av fallen har terminologen strukit under ordet i grundform, men inte alltid samma ord med en ändelse på slutet. Detta kan bero på att man glömt notera alla fall men även att vissa terminologer räknade böjningsformer som en term, även om den hade flera förekomster, då det inte var angivet något annat in instruktionerna. Ibland har även det motsatta skett, alltså att terminologen strukit under ordet med en ändelse på slutet men inte alltid ordet i grundform. (t.ex. *variseringen* ströks under men inte *varisering*).
- b) Den manuella metoden inkonsekvent då vissa terminologer hade tagit med vissa termer, medan andra inte ansåg dem vara termer eller eventuellt glömde dem. (t.ex. *patientens* är understruken i en artikeln men inte i en annan).
- c) De få terminologer som hade gett sig på att även stryka under de termerna som har frekvensen högre än 1, gick miste om många sådana termer (t.ex. *asymtomatisk bakteriuri* ströks under 2 gånger men förekom 4 gånger sammanlagt i just denna artikel).
- d) Vissa termer hittades aldrig trots att de förekom fler än en gång i texten. Det totala antalet på extra termer som hittades efter att den tredje granskningen avslutades blev 59 stycken. Diabetiker förekom 4 gånger i en artikel men ströks inte under alls.

I den första granskningen, på TNC, hittades 595 termer av sammanlagt 14 147 ord. Därefter gjordes en ytterligare granskning som ett andra steg. Alla artiklar granskades ännu en gång manuellt och det hela resulterade i att 654 termer hittades sammanlagt denna gång. Noterbart är att 474 av dessa 654 termer förekom endast en gång i texten och att 93 stycken av dessa 654 är böjningstermer. En annan sak som gjordes var att bestämma hur många gånger varje termkandidat förekom sammanlagt i datan.

Vid liknande studier (Pasanen, 2005) har man oftast utgått ifrån att den manuella metoden är så gott som 100-procentig, men det finns ett stort antal nackdelar med att göra termextraktionen manuellt och att det dessvärre uppstår en hel del problem på vägen. Dessutom har metoden tagit upp mot 160 timmar att slutföra och det brukar ta ännu längre tid, speciellt om man ska vara flera stycken och argumentera för och emot att någonting är en term.

5.2) RESULTAT FÖR DEN AUTOMATISKA UTVÄRDERINGEN

När det gäller den automatiska delen finns det 3 olika resultat att rapportera.

I 25 %-undersökningen hittades det 242 termer, varav 229 av dem var enordstermer.

I 50 %-undersökningen hittades det 742 termer, varav 680 av dem var enordstermer.

I 75 %-undersökningen hittades det 1 301 termer, varav 1 058 av dem var enordstermer.

Majoriteten verkar klart vara enordstermer i alla fallen. För att jämföra dessa resultat betraktar man termernas frekvens, förekomst och mönster och jämför deras värde i form av en tabell.

Först ut att redovisas är jämförelsen av termernas förekomst av termlängden, det vill säga, hur många ord som termerna består av. Auto 25 står för undersökningen med noise ration ställd på 25 %. Auto 50 står för undersökningen med noise ration ställd på 50 % och Auto 75 står för undersökningen med noise ration ställd på 75 %.

| X-ordstermer | Manuellt | Auto 25 | Auto 50 | Auto 75 |
|--------------|-------------|-------------|-------------|--------------|
| 1 | 72,9% (477) | 94,6% (229) | 91,6% (680) | 81,3% (1058) |
| 2 | 22,9% (150) | 3,3% (8) | 6,2% (46) | 12,1% (158) |
| 3 | 3,2% (21) | 0% (0) | 1,2% (9) | 3,6% (47) |
| 4 | 0,8% (5) | 1,7% (4) | 0,7% (5) | 0,9% (12) |
| 5 eller fler | 0,2% (1) | 0,4% (1) | 0,3% (2) | 2,1% (26) |
| TOTALT | 100% (654) | 100% (242) | 100% (742) | 100% (1301) |

Tabell A. Jämförelsen av resultaten för de olika metoderna

Resultatet av utvärderingen mäts med precision och täckning, beräknat i procent. Det går även att få fram ett medelvärde. Då använder man sig av F-score.

Precisionsmättet:

Precision = (antal korrekta extraherade dokument / antal extraherade dokument)

Precision beskriver hur stor andel av de extraherade dokumenten som är korrekta.

Täckningsmättet:

Täckning = (antal korrekta extraherade dokument / totalt antal korrekta dokument)

Täckning beskriver hur stor andel av det totala antalet korrekta dokument som har plockats ut.

F-scoremättet:

F-score = $2 * \text{Precision} * \text{Täckning} / (\text{Precision} + \text{Täckning})$

F-score räknar ut ett medelvärde av precision och täckning.

En redovisning av täckning, precision och F-score, av hela resultatet, i form av en tabell:

| Utvärdering | Täckning | Precision | F-score | Antal Giltiga Termer |
|-------------|----------|-----------|---------|----------------------|
| Auto 25 | 7,60% | 20,70% | 11,12% | 50 |
| Auto 50 | 22,50% | 19,80% | 21,06% | 147 |
| Auto 75 | 33,30% | 16,80% | 22,33% | 218 |

Tabell B. Jämförelsen av täckning, precision och F-score för de automatiska modellerna

En redovisning av täckning och precision, fördelat på 1, 2, 3, 4 och 5 eller flerordstermer, i form av en tabell:

| Utvärdering | X-ordstermer | Täckning | Precision | Antal giltiga termer |
|-------------|--------------|----------|-----------|----------------------|
| Auto 25 | | | | |
| | 1 | 9,20% | 19,20% | 44 |
| | 2 | 3,33% | 62,50% | 5 |
| | 3 | 0,00% | 0,00% | 0 |
| | 4 | 20,00% | 25,00% | 1 |
| | 5 eller fler | 0,00% | 0,00% | 0 |

| Utvärdering | X-ordstermer | Täckning | Precision | Antal giltiga termer |
|-------------|--------------|----------|-----------|----------------------|
| Auto 50 | | | | |
| | 1 | 27,88% | 19,56% | 133 |
| | 2 | 8,76% | 28,33% | 13 |
| | 3 | 0,00% | 0,00% | 0 |
| | 4 | 20,00% | 25,00% | 1 |
| | 5 eller fler | 0,00% | 0,00% | 0 |

| Utvärdering | X-ordstermer | Täckning | Precision | Antal giltiga termer |
|-------------|--------------|----------|-----------|----------------------|
| Auto 75 | | | | |
| | 1 | 39,41% | 17,77% | 188 |
| | 2 | 18,67% | 17,67% | 28 |
| | 3 | 4,76% | 2,13% | 1 |
| | 4 | 20,00% | 8,33% | 1 |
| | 5 eller fler | 0,00% | 0,00% | 0 |

Tabell C. Jämförelsen av täckning och precision för de giltiga termerna

När det gäller fördelningen över ordklasser var de flesta enordstermer och dessa var substantiv. När det gäller flerordstermer så var det mestadels nominalfraser som innehöll flera substantiv i rad eller av ordningen preposition + adjektiv + substantiv. I 95 % av fallen var det alltså antingen bara substantiv eller preposition + adjektiv + substantiv.

6) Diskussion

För det svenska språket resulterade utvärderingen med noise ratio inställt på skalan 25 %, 50 % och 75 % minsann på olika sätt. I vissa fall drabbades man starkt av övergenerering och i andra fall starkt av undergenerering. Baserat på många liknande tidigare undersökningar sägs det att nominalfraser bestående av två ord oftast omfattar mönstret av hur en godtagbar term ser ut. En ytterligare generell företeelse är att termer förekommer oftast fler än en gång i texten eller korpusen. Då hela 72,9 % av de svenska termerna från läkartidningar var enordstermer och 73,1 % av samtliga termer förekommer endast en gång i texten så kan denna generella företeelse, som sannerligen stämmer, orsaka problem och utesluta många giltiga termer. Likt Pasanens (2005) studie visar denna undersökning att adjektiv, verb och adverb ensamma väldigt sällan är giltiga termer. Ord med höga frekvenser och ett mönster som matchar (vilket är allt som oftast en nominalfras) fungerar utmärkt som en giltig term i de flesta fallen. Det finns för- och nackdelar med de tre utvärderingar som gjordes automatiskt men det finns även för- och nackdelar med att göra undersökningen manuellt.

6.1) DISKUSSION OM MANUELL UTVÄRDERING

Som vi har nämnt i resultatdelen gjordes den manuella utvärderingen i tre steg, inklusive allt efterarbete. Detta gjordes för att utvärderingen skulle bli så effektiv som möjligt. Dock tror jag aldrig att den blir fullständig, oavsett antal steg man tar och oavsett hur många gånger man korrigerar vissa detaljer i efterhand. Efter att första steget var genomfört var det 59 termer med och utan ordändelser som inte hade inkluderats som giltiga termer. En granskning gjordes då, som steg 2, men det är inte säkert att man fick med alla de termerna som saknades från steg 1. Vidare så stötte man på i steg 3 problemet med att människor är inkonsekventa. Frekvensen på hur många gånger en term förekommer i datan överensstämde inte och detta fick göras om. Förvisso kunde en del av dessa problem ha undvikits om instruktionerna hade varit tydligare, dock inte allting. För att visa några exempel på att människan är inkonsekvent markerades exempelvis inte ”diabetiker” trots att ordet förekom fyra gånger. Ett annat exempel är ”screening” som fanns med i tre olika artiklar men extraherades inte av mer än endast en terminolog. ”Patientens” var understruken i en artikel men inte i en annan.

Sedan kvarstår det största problemet i att veta om en term verkligen är en term och var termen i uttrycket, frasen och meningen börjar respektive slutar. Om vi tar oss an meningen ”En patient med arteriosklerotiskt betingad cirkulationsinsufficiens” så är det inte självklart vad som är termen. Människan valde ut ”arteriosklerotiskt betingad cirkulationsinsufficiens” medan Trados MultiTerm Extract tog ut adverbet ”arteriosklerotiskt” och substantivet ”cirkulationsinsufficiens” som två separata termer. Sanningen är att både människan och Trados MultiTerm Extract har rätt i det här fallet då det går att göra på båda sätten. Men i resultatet så kommer den automatiska att bokföras som fel då den inte överensstämmer med den manuella. Detta är bara ett exempel där det är svårt att avgöra vad i själva verket är en term. Där har vi ett stort problem.

6.2) DISKUSSION OM AUTOMATISK UTVÄRDERING

Av de tre automatiska undersökningarna som utvärderades i detta arbete så fungerar Auto 50-undersökningen bäst, i mitt tycke. Denna slutsats har jag dragit då Auto 50 drabbas minst av under- och övergenerering och har både en täckning och precision som är okej. Detta till skillnad från de övriga två automatiska undersökningarna där antingen täckningen eller precisionen är bra medan den andra delen verkligen är botten på grund av över- eller undergenerering av programmet. Men jag måste tillägga att även Auto 50 är långt ifrån toppen. Den största nackdelen med att använda den automatiska metoden är som sagt att det blir både övergenerering och undergenerering i många fall och att resultatet inte blir så lyckat vid större textmassa. Men det finns likväl fördelar med att använda sig av den automatiska metoden. När det gäller tiden tar det mindre än en minut att få ut resultatet av all data, vilket tar flera veckor med den manuella metoden att få fram. Dessutom är allting ordnat och ihopsamlat på ett strukturerat sätt i ett lätthanterligt program och allting är sparad i en dator som man senare kan enkelt redigera och använda sig av för vidare studier men även annan forskning. Verktöget Trados MultiTerm Extract lämnar dessvärre en del kvar att önska som faktiskt kan förbättras och vidareutvecklas i framtiden. Dessa idéer presenteras i nästa kapitel.

7) Avslutning

Även om det svenska språket inte är placerat på samma gren i det universala språkträdet som det finska språket så är de ganska lika varandra när det gäller frekvenser och förekomster av termer. Därmed stötte de även på samma slags problem. Språkträdet ser ut som följande:



Figur 4. Språkträdet (Söderberg, 2007)

I detta arbete har det utvärderats tre olika nivåer av MultiTerm Extract-verktyget. Den första, Auto 25, resulterade i en kraftig undergenerering och den sista, Auto 75, i en kraftig övergenerering. Auto 50 gav bäst resultat i och med att den drabbades minst av över- och undergenerering men även för att det totala antalet termer låg närmast det manuella resultatet. Resultatet i sig är väl inget lysande direkt men stora fördelar som att det inte kräver mycket tid att genomföra undersökningen och därigenom ta fram vissa giltiga termer är till stor fördel. Den manuella undersökningen är däremot så nära 100 % som man kan komma, men detta efter 3 stycken granskningar som skedde under sammanlagt en månads tid (160 timmar), efter alla redigeringar med cirka 10 personer involverade. Då ska man även notera att ingen expert från det specifika området deltog denna gång. Vill man åstadkomma bra manuellt resultat så är det tids- och resurskrävande. Vill man åstadkomma bra automatiskt resultat så är det inte så tidskrävande och inte resurskrävande alls då allting kan fullbordas av en person.

I det korta loppet, alltså, om man ska använda sig av att ta ut termer någon enstaka gång, så vinner den manuella metoden. Men i det långa loppet så är jag övertygad om att den automatiska metoden är värt att satsa på, just för att man ska spara in tid, pengar och resurser med denna metod. Dock ska det till en del förbättringar.

7.1) VIDAREUTVECKLING

Trados MultiTerm Extract är ett fungerande verktyg med stor potential att vidareutvecklas. Programmet är lätt att använda och består inte av några som helst komplikationer. Det ingår inte många steg för att exekvera det hela och få ut ett välstrukturerat och användarvänligt resultat. Programmet innehåller flera nyttiga funktioner som är användbara i det långa loppet. Så detta är definitivt ett verktyg som är värt att satsa på. Dock finns det rum för förbättring. För att åstadkomma bättre resultat, vilket ska prioriteras med tanke på vad resultatet blev utav denna undersökning, så kommer jag här med några personliga förbättringsförslag som förhoppningsvis är värda en tanke.

I språkträdet i figur 4 såg vi att trots att svenska och finska inte befinner sig på samma gren så blev resultaten ganska så lika varandra och man stötte på samma slags problem. Jag är övertygad och håller med om Pasanen (2005) som tror att resultatet blir mycket bättre för de språk som har en termdatabas. Det första som behöver göras i förbättringsväg är att skapa en sådan bas för de övriga språken också. Att skapa en termdatabas är arbetskrävande men man har hjälpmedel. Exempelvis kan man använda sig av dataprogram för att göra en termdatabas och sedan lägga termer, definitioner och all annan viktig information i databasen. Dataprogrammet organiserar olika fält så att databasen ser bra ut. En termdatabas bör därmed skapas för svenska och helst med den funktionen att man även manuellt sedan på ett enkelt sätt ska kunna hämta in och spara nya termer som känns igen vid nästa exekvering av nya texter.

Vidare vore det bra om man kunde programmera verktyget så att allt som man har gjort innan kan omanvändas, det vill säga, att informationen sparas i programmet, typ i en termdatabas eller liknande. I nuläget startas projektet helt från början och använder sig inte av någon slags information eller minne från de tidigare körningarna. Därmed är något slags systemminne ett måste för det långa loppet.

MultiTerm Extract använder sig utav en stoppordlista på 306 ord för det svenska språket. Denna lista innehåller vanligast förekommande icke-termer och denna lista med litet antal ord måste bara utökas med flera icke-termer. Förutom att fylla på med nya ord så måste även varianterna och synonymerna av dessa inputs också inkluderas. Så nästa storsatsning på vidareutveckling måste komma här om man ska kunna i fortsättningen på ett smidigt sätt lägga in icke-termer i denna lista så fort man upptäcker sådana, troligtvis, efter att ha kört en automatisk termextraktion.

Ytterligare en sak som kan förbättras är att verktyget inte kan läsa in bland annat PDF-filer, vilket är väldigt synd då det mesta på akademisk nivå publiceras just i PDF format. Detta skulle spara den tid det tar att kopiera texterna och klistra in dem i ett annat fungerande format för verktyget. När man kopierar och klistrar in från ett format till ett annat löper man vissa risker såsom att något blir bortglömt men även att något ibland inte kommer med.

Allt i allt, det ultimata som borde göras, i mitt tycke, för att förbättra Trados MultiTerm Extract radikalt för det svenska språket, skulle vara att först och främst skapa en termdatabas i vilken man kan spara termer, även manuellt, vilket säkerligen behövs och ha igång denna termdatabas för det svenska språket vid varje körning av en svensk text. Därefter ska man även ha en stoppordlista som man också kan korrigera manuellt genom att fylla på med icke-termer och dess varianter, allteftersom man stöter på nya sådana. När det gäller att läsa mönster sköter programmet det bra.

Jag är övertygad om att resultatet skulle förbättras ifall man startade med dessa åtgärder. Därmed drar jag slutsatsen att det är värt att använda sig utav termextraktionsverktyget Trados MultiTerm Extract, men att man måste utveckla det ett par steg till för att de ska kunna prestera åtminstone 50 % när det gäller precision och täckning.

Litteraturförteckning

Alegria, I., Gurrutxaga, A., Lizaso, P., Saralegi, X., Ugartetxea, S., och Urizar (2004). ”R. Linguistic and statistical approaches to Basque term extraction”.
URL <http://citeseer.ist.psu.edu/650853.html>

Arppe, Antti (1995). “Term extraction from unrestricted text. I: NODALIDA-95”.
URL <http://www2.lingsoft.fi/doc/nptool/term-extraction.html>

Bowker, Lynne (2003). "Terminology tools for translators". I: Somers, Harold (ed.) (2003), s. 49–65.

Church, Kenneth & Hanks, Patrick (1989). “Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics” Vancouver, B.C., Association for Computational Linguistics.
URL <http://citeseer.ist.psu.edu/church89word.htm>

Esselink, Bert (2003). "Localisation and translation". Ur: Somers, Harold (ed.) (2003), s. 67–86.

Eugen Wüster (1931). ”Internationale Sprachnormung in der Technik: Besonders in der Elektrotechnik”, Sprachforum. Beiheft.

Granitzer, Michael (2003). “Classification of hierarchical document spaces using machine learning technologies”. Doktorsavhandling, Graz University of Technology, Institute for Theoretical Computer Science (IGI).

ISO 10241 (1992). *International terminology standards – Preparation and layout*.

Justeson, J.S. & Katz, S.M (1995) “Technical terminology: Some linguistic properties and an algorithm for identification in text” Ur: Natural Language Engineering.

Manning, Christopher D. och Schütze, Hinrich (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

NaviTerm 2.0 (2007). URL <http://www.naviterm.com>

Nordterm (2005). *Terminologins terminologi*.

Pasanen, Päivi (2005). “A term list or a noise list? How helpful is term extraction software when Finnish terms are concerned?” I: *7th International Conference on Terminology and Knowledge Engineering*, ss 375–383, Köpenhamn, Danmark, 2005.

SDL MultiTerm Extract (2007). URL <http://www.sdl.com/products-home/products-home/products-sdlmultiterm-extract.htm>

SDL Phrase Finder (2005). URL <http://www.translationzone.com/en/products/sdlphrasefinder/>

Soininen, Pirjo (1998). ”Terminhaun automatisointi”. URL <http://www.ling.helsinki.fi/~psoinine/lp.html>

Somers, Harold (ed.) (2003). “*Computers and Translation: A Translator's guide*”. Amsterdam/Philadelphia: John Benjamins.

Suonuuti, Heidi (2004). ”*Terminologiguiden. En introduktion till terminologiarbete i teori och praktik.*” Terminologikum TNC.

System Quirk (2007). URL <http://www.computing.surrey.ac.uk/SystemQ>

Söderberg, Boel (2007). *Pärm 2: Språkhistoria, ordkunskap, uppsats och diktanalys*. Ekelunds förlag.

SPRI, rapport 481 (1999) *Metoder och riktlinjer för terminologiarbete*. Hälso- och sjukvårdens utbildningsinstitut. Stockholm: Spri.

The Pavel Terminology Tutorial (2006). URL http://www.termium.gc.ca/didacticiel_tutorial/english/lesson1/index_e.html

TNC och Spri (1999). *Metoder och principer i terminologiarbetet*. URL <http://www.sos.se/epc/klassifi/filer/Terminologi/rap481.pdf>. Spri rapport 481.

TNC (2007). *Terminologilärans grunder*. URL http://www.tnc.se/index.php?option=com_content&task=view&id=104&Itemid=93

User Guide MultiTerm 7 Extract (2005). Trados

Vintar, Špela (2005) *Študijska gradiva*. URL <http://www2.arnes.si/~svinta/terminologija02-03.doc>

Wihuri, Paavo (2002). *Meriliikenteen ohjausjärjestelmät, VTS ja VTMISS*. A seminar paper presented in Uusikaupunki.

Wikipedia (2007) URL <http://sv.wikipedia.org/wiki/Terminologi>