GÖTEBORGS UNIVERSITET

Autentiska och artificiella frågor till svensk text

Automatisk frågegenerering jämfört med användares frågor för informationsåtkomst

Authentic and artificial questions to Swedish text

Automatically generated questions versus user-generated questions for information access

Kenneth Wilhelmsson

Kandidatuppsats i informatik

Rapport nr. 2015:064
ISSN: 1651-47

Göteborgs universitet
Institutionen för tillämpad informationsteknologi
Göteborg, Sverige, 2015

Tack

Jag vill tacka *Marie Eneman, Magnus Bergqvist* och *Arvid Birkenmeier Selin* för synpunkter och rekommendationer gällande litteratur.

Jag vill också tacka de anonyma undersökningsdeltagarna som avsatte tid åt uppgiften.

Abstract

Information access using unstructured data sources like free text is one of the areas where natural language user interfaces have been introduced. In such a (possibly *AI*-oriented) system, a few basic difficulties can be noted. One such difficulty emerges from the fact that a user is unaware of whether a particular question to be posed is in fact answered by the current text database. These difficulties, together with other problems, like the great variation of linguistics expressions that an answering segment may come in, put the user experience of this type of system interface at risk.

The processes involved in such a *question answering system (QA)* must somehow incorporate a mapping from *wh*-word (or similar), like *when*, to the syntactic form and function of the plausible answer (for *when*, a temporal adverbial would be a likely candidate). These and other observations suggest that *question generation (QG)* might be a well-suited supporting technology. Question generation is a process of initial generation of questions which are answered by the natural text in explicit form. The idea of bringing this mechanism into the setting of information access means restricting the user of the system's user interface to only allow her to pose one of those questions, which do have answers.

This study deals with the questions that an automatic QG system for Swedish is, or, through further development, would be able to generate for arbitrary digital text in Swedish. Even though the amount of questions (and *reformulations*) may become very large, several times larger than the source text, it is clear that those sets do not, and probably will not, contain all questions – and formulations – that a human user would state that a certain text provides answers to. So, how well does automatically produced questions work for this task?

This thesis revolves around a user-study where the participants were asked to formulate relevant questions that texts answer. The resulting set of questions were examined and categorized. The result of the main question was that only about 20-25 % of the questions (formulations of questions) produced by the user could be generated automatically with the current technique for Swedish – without certain improvements on the generation side.

The study presents some new terminology (in Swedish) for coping with the varying degrees of technical improvements needed for production of different question types.

This thesis is written in Swedish.

Abstract

Informationssökning mot ostrukturerade datakällor som fri text är ett av de områden där användargränssnitt med fri formulering i naturligt språk har tagits fram. I ett sådant, eventuellt AI-betonat, system kan några grundläggande svårigheter från användarperspektivet märkas. En sådan svårighet är att en användare inte känner till huruvida en fråga som hon avser att ställa egentligen kan besvaras av den aktuella texten. Denna svårighet, tillsammans med andra, som de kraftiga variationsmöjligheterna för formen för ett giltigt svar på en ställd fråga, riskerar att leda till att användarintrycken av systemtypen blir negativa.

De moment som behöver ingå i ett sådant frågebaserat informationssystems funktionssätt måste på något sätt inbegripa en mappning av frågeled i frågan (t.ex. *när*) till den form och grammatisk funktion som svaret i texten måste ha (för frågan *när* normalt ett tidsadverbial). Bland annat denna iakttagelse inbjuder till användning av *automatisk frågegenerering (question generation, QG)*. Frågegenerering innebär att frågor som en naturlig text besvarar initialt utvinns av ett program som samlar in dem i explicit form. Tanken för användning i informationssökning är att en användare i gränssnittet enbart ska kunna ställa just dessa frågor, vilka faktiskt besvaras av texten.

Denna studie gäller just de frågor som ett automatiskt frågegenereringssystem för svenska kan, och genom vidare utveckling, skulle kunna generera för godtycklig digital svensk text. Även om mängden automatiskt genererade frågor och frågeformuleringar kan bli mycket stor, utrymmesmässigt många gånger större än ursprungstexten, så är det tydligt att den beskrivna metoden för frågegenerering för svenska inte kan och troligen inte heller kommer att kunna förmås att skapa alla de frågor och frågeformuleringar som en vanlig användare skulle anse att en viss text besvarar. Men hur väl fungerar då automatiskt genererade frågor i detta sammanhang?

Denna uppsats kretsar kring en användarundersökning där undersökningsdeltagare har ombetts att formulera frågor som texter besvarar, och som anses vara relevanta frågor. Den resulterande samlingen frågor undersöktes och kategoriserades. Resultatet av undersökningens huvudfråga visar att bara 20-25 % av användarnas frågeformuleringar skulle kunna genereras direkt automatiskt med aktuell ansats – utan vissa informationstekniska förbättringar.

Uppsatsen föreslår viss ny terminologi för detta outforskade område, bl.a. för att skilja mellan de olika grader av processkrav som generering av olika frågeslag från text kräver.

Innehåll

1 Inledning

Begreppet *information* förekommer i litteraturen med flera olika definitioner, som skiljer sig åt beroende på sammanhang, se t.ex. Beynon-Davies (2013) eller Ribeiro-Neto (1999). När informationsformen är skriftliga textdokument förekommer i vardagligt tal en lekmannamässig definition av information: *en texts information är de frågor som den besvarar*.[1] Om denna definition skulle få råda mer allmänt, så skulle förmodligen dessa tillhörande, besvarade frågor ges ett större fokus, och inbjuda till systematiska studier.

En faktatyngd textsekvens som den nedanstående är informationsrik ur de flesta synvinklar. Texten är idealisk som frågebesvarande material för det ändamål som avses här. Detta arbete handlar på ett sätt om den stora skillnaden mellan att ställa fråga a, b eller c nedan mot texten genom ett söksystem med naturligt språkgränssnitt.

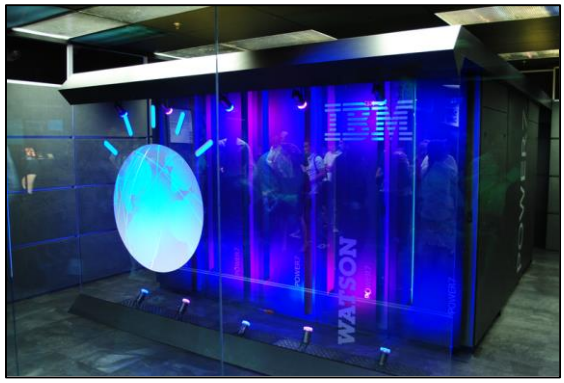
> Visby har 22 593 invånare. Bland de mest anmärkningsvärda historiska lämningarna är den 3,4 km långa ringmuren som omger staden och dess gamla kyrkoruiner. Visby är ett populärt resmål under sommaren och tar emot tusentals turister varje år. Visby är säte för Högskolan på Gotland. Visby kallas 'rosornas och ruinernas stad'.
>
> Från *Visby* (Svenska Wikipedia), modifierat.

a) Vad kallas Visby?
b) Hur många bor i Visby?
c) Är Visbys ringmur längre än en mil?

1.1 Bakgrund: Frågebesvarande system för informationssökning

Under de senaste åren har systemtypen *Natural Language Query Systems*, söksystem med naturligt språkgränssnitt, varit på frammarsch. Speciellt stor uppmärksamhet fick en implementation för engelska (”*Watson”*)[2] som visade sig framgångsrikt kunna besvara frågor i en TV-sänd frågesport genom att snabbt använda naturlig text och extrahera rätt segment för det aktuella informationsbehovet. Det har funnits flera system av denna typ, t.ex. *PowerSet* (Converse et al 2008), och *Ask Jeeves*,[3] vilka gemensamt kan kallas *frågebesvarande system (QA systems)*. Somliga har varit fritt tillgängliga på Internet (se vidare nedan).

[1] I engelska *Wikipedia* är definitionen av *Question*: “A *question* is a linguistic expression used to make a request for information, or the request made using such an expression. The information requested may be provided in the form of an answer […]”
Samma källa inleder artikeln *Information*: ”Information (shortened as info or info.) is that which informs, i.e. an answer to a question, as well as that from which knowledge and data can be derived”. (Egna markeringar, kontrollerade källor 20150424.)
[2] (This is Watson (special issue), 2012)
[3] www.ask.com



Figur 1: Datorsystemet *Watson*.[4]

Det fokus på konstruktion och användning av anpassade informationssystem som finns inom informatik förutsätter oftast att aktuell data finns i *strukturerad* form. Denna term – i sin vidaste bemärkelse – inbegriper genomarbetade relationsdatabaser eller objektorienterade databaser och platta såväl som hierarkiska datalagringsformat. Vanliga tekniska format för strukturerad data i industrin idag inkluderar generella format som XML och JSON. Över huvud taget förekommer data som ska behandlas effektivt i *någon* strukturerad form.

Digital-iserat	*Webb*	*Databas*
Icke-Digitaliserat	*Konversation*	*Bibliotek*
	Ostrukturerat	**Strukturerat**

Tabell 1, Typuppdelning hämtad från Stenmark (2002), s. 8, utgående från Davenport (1997).

Textdokument skrivna för mänskliga läsare är inte databaser – åtminstone inte i denna betydelse. Inte desto mindre är det en återkommande förhoppning att texter skulle kunna behandlas ungefär som om de vore databaser. Naturlig text, som i tekniskt sammanhang kan kallas *ostrukturerad*, hanteras ideligen i informationssystem, speciellt för uppgiften *sökning*, med processer som inte alls sker i enlighet med skriftspråkets primära funktion – dvs. att bli läst och förstås av mänskliga läsare. Det finns enorma mängder av information uttryckt i naturligt språk. Det går att söka efter informationen men jämfört med strukturerad data är den mindre hanterbar. Går det då att öppna upp denna information för mer raffinerad datoriserad hantering?

Hur ska systemgränssnitt med naturligt språk-gränssnitt som *Watson* ovan klassificeras? Rogers et al. (2013, s 215-217) använder termen *natural user interfaces (NUI)* som en samlingsterm för andra sätt att interagera med ett implementerat informationssystem än de förhärskande grafiska användargränssnitten (GUI). NUI är en bred term som inbegriper interaktion inte bara med hjälp

[4] Bildkälla: Wikipedia, GNU Free Documentation License.

av naturligt språk, t.ex. röststyrning, utan även gester och *touch*. Rogers et al. (2013) utvecklar ett sunt, lite ifrågasättande perspektiv på dessa moderna former av användardesign. Den centrala termen ’naturlig’ (eller ’intuitiv’) kan många gånger kritiseras – vad är t.ex. den mest *intuitiva* kroppsliga gesten för att höja temperaturen (*ibid*. s 216) med hjälp av ett datorsystem? Vanans makt är stor. En kvalitet hos många gränssnitt är helt enkelt likhet med andra gränssnitt som är bekanta för användare (även om de inte nödvändigtvis är *naturligast* i någon absolut mening). Men vad liknar i så fall de nya typerna av gränssnitt?

I detta arbete finns användningsaspekter av en prototyp till ett informationssökningssystem med naturligt språkgränssnitt i huvudfokus. Den slutform av system som skisseras är ett program i vilket användaren ställer en vanlig naturlig fråga och får korrekt svar om det finns ett sådant i en textmängd. Det är emellertid (i detta läge) fråga om ett program med relativt traditionellt grafiskt användargränssnitt som används med tangentbord och datormus för input. Den aspekt som, enligt Rogers et al. (2013), skulle göra det till ett NUI-system, är enbart det faktum att sökningen sker med naturligt språk, här svenska. Det är emellertid skriven digital text som används i gränssnittet, inte t.ex. talgränssnitt. Det betyder att systemet här har ett regelrätt grafiskt gränssnitt även om det genom naturligt språk i användningen bär drag av NUI.

1.2 Ett problem med QA-system och en förutsättning för uppgiften

Enligt begreppsapparaten hos Beynon-Davies (2013) hamnar den aktuella studien i något som där benämns gränssnittslagret (*interface layer*, s. 24). Även den konkreta gränssnittstypen, *natural language interfaces (NLI, naturligt språkgränssnitt)*, nämns av författaren. Beynon-Davies skriver hur de NLI-system han åsyftar (oklart vilka) har en svaghet genom att de enbart fungerar med vissa formuleringar; *Give me all the salaries of all my employees* ger kanske bra resultat medan *List my employees' salaries* inte gör det, s. 279.

För *frågesystem med naturligt språk*, som exemplifierades ovan, vilka söker sina svar i samlingar av naturlig text, ska här ett nödvändigt steg i processen, och en inneboende svaghet hos själva systemtypen inledningsvis tas upp. De två kommande avsnitten i detta kapitel syftar till att belysa två centrala sidor hos sådana system, varav ett är uppsatsens *problem* – samt ett möjligt sätt att eventuellt råda bot på det.

- En karakteristisk egenskap i själva förutsättningarna hos frågebesvarande system: *vilka former kan ett svar till en viss sorts fråga ha, och vad får det för konsekvenser?*

- En allmän inneboende svårighet för användaren av ett frågebesvarande informationssystem som tar emot en naturligt formulerad fråga och finner svaret, i naturligt språk: *Hur vet en användare av ett informationssystem (oavsett typ) ifall en fråga egentligen kan besvaras utifrån den tillgängliga databasen? Och fungerar hennes naturliga frågeuttryck rent formuleringsmässigt?*

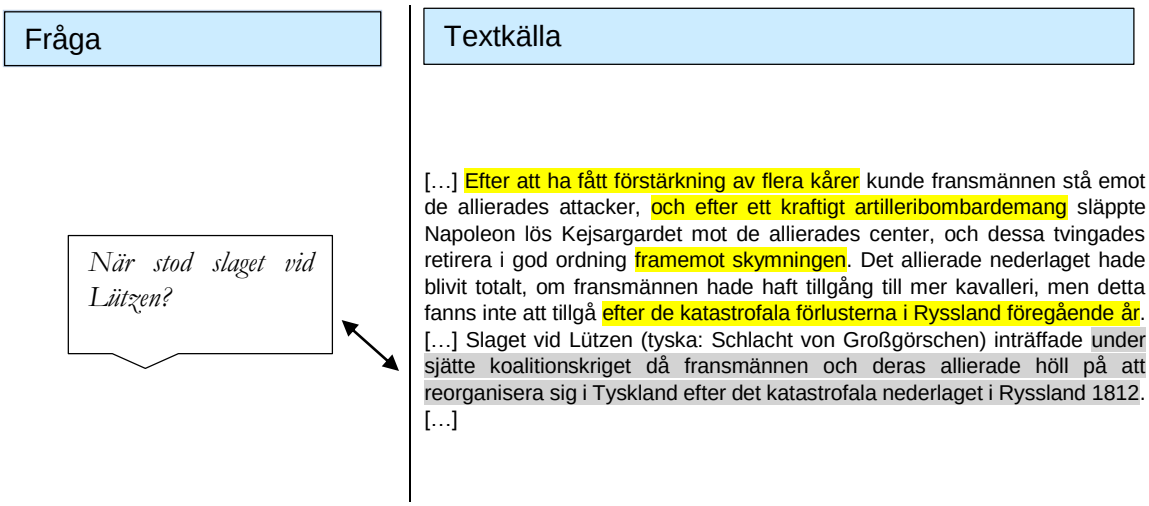
Den huvudsakliga *frågeställningen* i detta arbete utgår från dessa svårigheter vid användning av ett informationssystem där användaren ställer frågor mot en textdatabas. Närmare bestämt är *syftet* att undersöka om en speciell, idag delvis implementerad teknik för svenska, verkar ha användbara egenskaper i sammanhanget – en teknik som förprocessar texten som är informationskällan och *samlar in* frågor som den besvarar. – I vilken mån kan en sådan process bemöta ett verkligt informationsbehov? I vilken grad är de frågor som då skapas sådana som en användare skulle anse vara *relevanta?*

Metoden är en användarundersökning som försöker utröna ungefär *hur stor andel av frågor som verkliga användare anser är relevanta utifrån en text, som skulle kunna genereras automatiskt,* och därmed byggas in i systemet, för att underlätta informationssökningsprocessen. I praktisk användning blir det dessutom fråga om precis vilka *formuleringar* av frågor (dvs. just ”formuleringsaspekten” ovan, enligt Beynon-Davies, 2013, s. 279) som skulle kunna genereras.

Detta är emellertid att gå händelserna i förväg. Innan det blir helt uppenbart varför den nämnda tekniken, dvs. att samla in besvarade frågor, över huvud taget skulle kunna ha en relevant roll i ett naturligt språkgränssnitt för informationssökning behövs här mer bakgrund. Problem, syfte och frågeställning utvecklas vidare i de följande avsnitten.

1.3 Relationen mellan frågetyp och frågebesvarande led

För att ett informationssystem ska kunna ge ett svar på en fråga som ställts i naturligt språk krävs oavsett ansats några fundamentala procedurer hos metoden. En avgörande sådan procedur är att kunna bestämma vilken form ett svar på en fråga kan ha. En fråga om en tidpunkt, som en *när*-fråga, kräver sålunda att textsegmentet som ska utgöra svaret är ett tidsuttryck. En *vem*-fråga kräver på samma sätt att svaret har formen av en *nominalfras (NP),*[5] närmare bestämt att det har en *animat* (person, djur eller organisation) referent. En *varför*-fråga kan i de enklaste fallen ha svar på formen adverbial – t.ex. en *eftersom*- eller *pga*-bisats. Men potentiella svar på just en sådan fråga (en *varför*-fråga) finns även på andra håll, däribland i potentiellt i långa sjok av textmeningar som måste kombineras enligt logiska regler (se vidare i 2.6). Sett i detta perspektiv blir det tydligt att naturliga frågor ofta kan ha många olika möjliga svarsformer och att själva dessa, eventuellt besvarande, strukturtyper kan finnas på väldigt många ställen i en större textsamling som fungerar som databas.



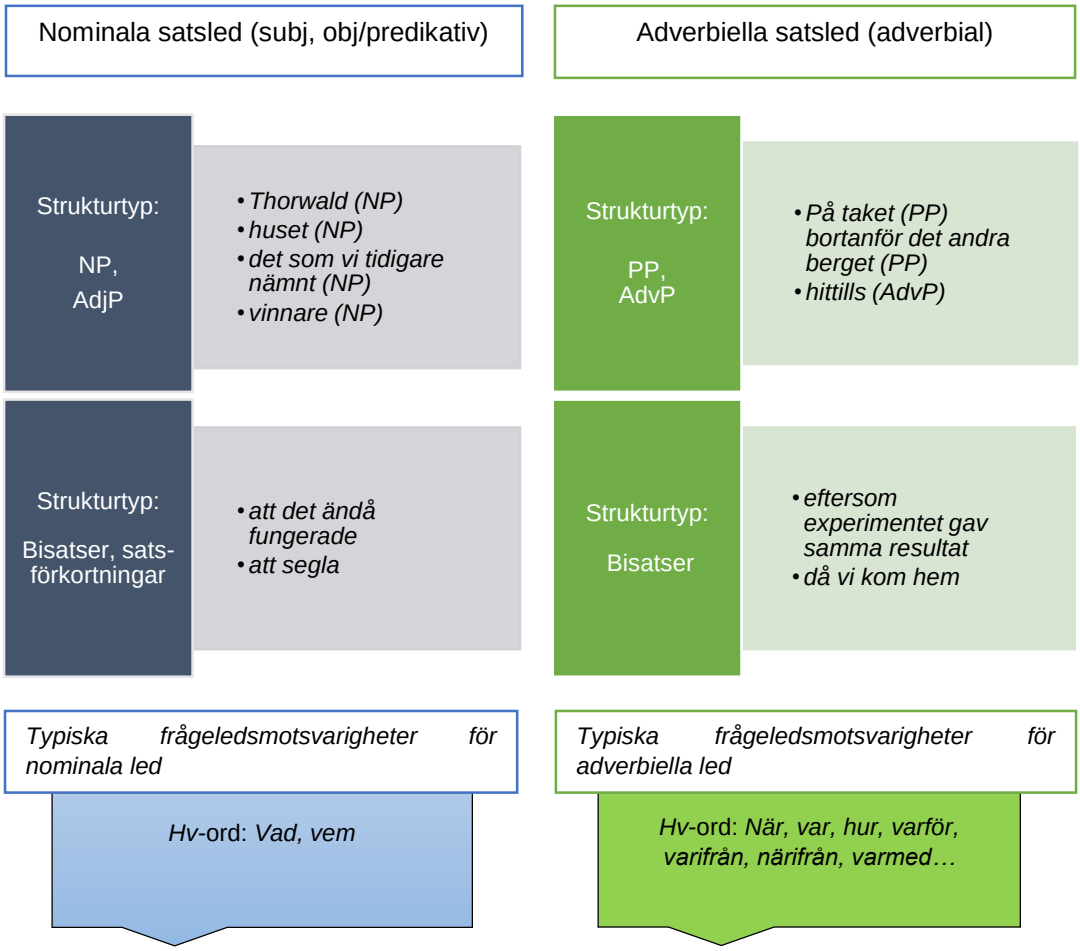
Figur 2 I ovanstående exempeltext finns tidsadverbial markerade. Det sista tidsadverbialet utgör det relevanta svaret på den aktuella frågan. Text hämtad från Wikipedia (modifierad).[6]

Är texten tillförlitlig? (Det kan vara värt att notera att texten, och därmed svaret, i exemplet ovan gäller det andra, fransk-tyska, slaget vid Lützen 1813 – det är inte en oväsentlig aspekt i detta

[5] En förenklad (och haltande) definition av NP är att det är de segment som kan vara subjekt eller objekt/predikativ.
[6] Kontrollerad 150909.

sammanhang att somliga begrepp svarar mot flera olika företeelser.) I de flesta fall besvaras en viss fråga av de typer som här undersöks bara (högst) en gång av en enskild text men det finns fall då identisk information upprepas i en faktatext. När det gäller skönlitterära texter kan svar på en viss fråga finnas på olika håll – och dessutom vara olika. Detta beror ibland på att texten skildrar ett tidsförlopp där giltig fakta i berättelsen förändras under tidens gång. Således kan olika delar i texten ange (svara på en fråga) att det är sommar respektive vinter, att två personer inte har träffats, respektive att de har träffats osv. (Wilhemsson, 2012).

Förutom dessa förhållanden finns fler intressanta aspekter rörande hur *relationen mellan frågetyper och de möjliga svaren* ser ut – hur fungerar grammatiska led i text som besvarande segment? För att klargöra nedanstående resonemang behövs en smula grammatisk terminologi. En *funktionell* grammatisk analys är här en uppdelning i *adverbiella* led (de som är adverbial) och *nominala* led (sådana som är subjekt eller objekt/predikativ). En *strukturell* grammatisk beskrivning, å andra sidan, inbegriper syntaktiska *frastyper*. Det är det *funktionella* synsättet, först och främst en uppdelning mellan adverbiala och nominala led, som kommer att ha störst betydelse här i frågesammanhanget. Ett visst funktionellt led som *adverbial*, och mer bestämt t.ex. ett tidsadverbial kan sedan *strukturellt* sett vara av flera olika *strukturtyper*, t.ex. PP (prepositionsfraser, vilka inleds med preposition), eller AdvP (adverbfras), se nedan.



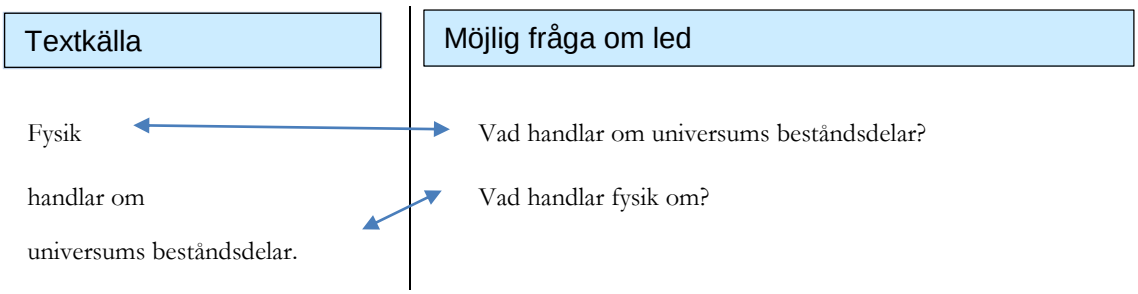
Figur 3: Relationen mellan funktionella ledslag och hv-frågeord för svenska är en viktig startpunkt för utröna frågors möjliga svarspositioner i en text. (De nominala och de adverbiella satsleden har vissa typiska strukturformer vilka visas i figuren.) Nominala led omfrågas generellt med en mycket mindre uppsättning frågeord (vad/vem/vilka) än de adverbiella leden.[7]

En uppdelning av texters ingående huvudsatser i *adverbiella* (adverbialled) och *nominala led* (de som kan utgöra subjekt och objektiv/predikativ), som i figuren ovan, pekar alltså på att en viss frågetyp, exempelvis en *var*-fråga skulle kunna besvaras med många olika former (främst tidsadverbial) som strukturellt kan vara prepositionsfraser (*till*-PP, *i*-PP, *vid*-PP, *under*-PP etc.), bisatser: (*där*-bisatser osv) och adverb-fraser: (med huvudord som *där, hitåt* etc.). Se vidare nedan.

Vad är då poängen med uppställningen ovan? Iakttagelsen som är värd att göra här är att fastän *svaret* på en naturlig fråga alltså kan ha många olika strukturformer (särskilt de adverbiella frågorna) – så är inte relationen i motsatt riktning likadan i detta avseende: Varje textsegment, dvs. ’svaren’, tillhör i allmänhet bara en viss informationstyp, de besvarar i allmänhet bara en fråga/frågetyp var.[8]

[7] Även val av frågeled för de nominala leden kräver i många fall en mer raffinerad analys för att välja rätt *hv*-led.
[8] Det liknar därmed en *ett-till-ett*-relation mellan led och frågor, vilket skulle innebära att kraven uppfylls för att vara en giltig matematisk funktion. Det finns emellertid komplikationer som att led inte tydligt besvarar någon fråga alls.

Exempelvis är rumsadverbialet *I Washington* något som enbart kan besvara en *var*-fråga – samtidigt som varje *var*-fråga alltså skulle kunna ha många andra svarsstrukturer (t.ex. en adverbiell bisats som *där hammaren ligger*).[9]



Figur 4: Om betraktelsen utgår från texten (’svaren’) istället framgår att varje informationssegment är ett svar på en enda fråga (något förenklat uttryckt).[10]

Resonemanget ovan och figur Y syftar till att klargöra att om texten (och i det sönderdelade perspektivet: dess huvudsatsers funktionella led) istället får vara utgångspunkten och betraktas ungefär som ’en mängd svar’, så kan själva uppgiften för frågebesvarande system framstå som mer överskådlig och möjliggöra ett mer systematiskt tillvägagångssätt. Den önskade poängen här är att det kan vara rationellt att utgå från själva texterna på ett systematiskt sätt – istället för att utgå från frågorna och låta systemet ’leta efter möjliga svar’.

1.4 Problem – En svårighet för frågebesvarande system: Användaren och systemet vet inte om frågan kan besvaras

I ett metodologiskt perspektiv går det att analysera frågebesvarande informationssystem för naturligt språk som byggts som de som nämndes i inledningen, och säga något om deras allmänna begränsningar i ett användarperspektiv. För det första gäller det hur informationsbehovet i form av en naturligt uttryckt fråga behandlas.

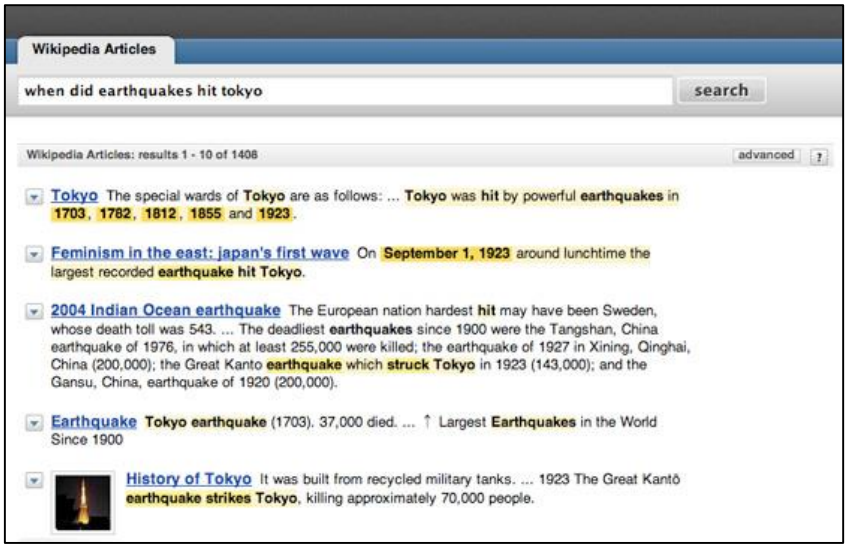
Istället för vad som kanske vore önskvärt – en relativt djup analys från systemets sida – som i primitivare mening ’förstod’ både frågan och textinnehållet och kunde avgöra att något verkligen innebar ett korrekt svar på en viss fråga, fungerar flera system som besvarar naturliga frågor med fri text med en kombination regler och statistik. I fallet *PowerSet* används en kombination av lagrade semantiska relationer och statistik (Converse et al, 2008) – bland annat enkel strängmatchning av *n-gram*[11] – för att finna de antaget relevanta segmenten för den naturligt uttryckta frågan.

---

[9] För att komplicera detta något kan ett motexempel från litteraturen nämnas: En fråga i t.ex. ett tänkt olympiskt sammanhang: – *När var detta? – I Moskva.*

[10] För att inte skymma poängen med exemplet visas inte samtliga frågor (och speciellt inte omformuleringar av dessa).

[11] Ett *n-gram* är en sekvens av löpord av längden *n*: Exempelvis är bigrammen i *Malta är Europas sydligaste stat och en av Europas mikrostater* de följande: 'Malta är', 'är Europas', 'Europas sydligaste', 'sydligaste stat', 'stat och', 'och en', 'en av', 'av Europas', 'Europas mikrostater'. För ordklasstaggning med den den sannolikhetsbaserade statistiska modellen HMM med Viterbi-algoritmen (Viterbi, 1967), samlas bl.a. textdata in från naturlig och uppmärkt text, delvis i form av uni-, bi-, och tri-gram.



Figur 5: *PowerSet* i användning. [12]

Det frågebesvarande systemet *PowerSet* som använde Wikipedia som textkälla var just ett av de system som rönte uppmärksamhet för några år sedan.[13] Det fanns under en period fritt tillgängligt på Internet och tillät naturliga frågor på engelska. Som svar på en fritt formulerad fråga erhöll användaren (frågeställaren) en rankad lista av textsegment som svar på informationsbehovet. Textsegmenten visades alltså i den ordning som modellen beräknade som mest sannolikt frågebesvarande. Just där finns den nämnda svårigheten hos denna typ av system. Systemtypen analyserar ju inte text ”i enlighet med dess primära funktion”, att bli läst av människor), utan med ett mer sökliknande angreppssätt: detta betyder att systemet egentligen inte vet om de faktiskt har hittat svaret. – För att ytterligare förtydliga denna situation: systemet kan inte säga om det efterfrågade svaret alls finns i den aktuella textdatabasen. En probabilistisk ansats fungerar från början med en sorts inbyggd osäkerhet. I värsta fall är systemtypen inte så olik ordinär strängbaserad sökning och levererar blint vad som matchar bäst enligt algoritmen.

För att pröva denna svaghet hos metoden och ådagalägga komplikationen ställdes frågor som egentligen saknade svar, t.ex. *Who is the tallest Dane?* mot tjänsten PowerSet (Wilhelmsson, 2010). Resultatet pekade på en svaghet med statistiska ansatser. Det resulterade som alltid i en rankad lista av textsegment som enligt systemets funktionssätt var de bästa kandidaterna, men något giltigt svar fanns där förstås inte på just denna fråga.

Om ett informationssystem inte gör någon form av djupare analys av texten alls för att t.ex. identifiera grammatiska funktioner, dvs. *parsning* (se nedan) saknar systemet det mest grundläggande steget mot vad som ibland, lite vanvördigt, har kallats ’förståelse’ av texten.

1.5 En strategi för att bemöta fråga-svars-relationen och det inherenta problemet hos systemtypen

Det som nu karakteriserats hos frågebesvarande informationssystem mot naturlig text i de två avsnitten ovan kan sammanfattas på följande sätt.

[12] Bildkälla: Techcrunch.com, kontrollerad i augusti 2015.
[13] 2008 förvärvades PowerSet av Microsoft.

- En enskild fråga (frågeformulering) kan i många fall besvaras med olika former. Vilka segment i en text är då potentiella svar på en särskild fråga? *Kortfattat uttryckt kan svaret finnas i väldigt många former*. – Om istället själva texten får vara utgångsläge framträder hur varje informationssegment oftast svarar mot en viss fråga: detta inbjuder till ett systematiskt tillvägagångssätt.

- *Problemet:* De frågebesvarande informationssystem av hittills nämnda slag är inte kapabla att avgöra *om ett svarssegment som systemet föreslår helt säkert besvarar en ställd fråga*. Användaren av ett sådant system vet inte om svaret för ett visst informationsbehov tillgodoses i de texter som används. Användaren kommer kanske också till snabb insikt om att inte heller systemet kan avgöra det.

Dessa två förhållanden kan leda till följande idé: systemets kunskapsdatabas som ska innehålla användbara svar för att tillgodose en frågeställares informationsbehov ”består av en mängd möjliga svarsled” som i allmänhet besvarar en frågetyp var. Det skulle vara användbart att samla dessa potentiella frågesvar i en inledande process för att kunna se (åtminstone delvis) vilka frågor som texten över huvud taget kan tänkas vara kapabel att besvara.

– Men det går också att ta detta ett steg längre: genom syntaktiska processer (se nästkommande kapitel) är det möjligt att utifrån dessa möjliga svarsled generera frågor som besvaras. Det som här skisserats är alltså en idé som innebär att uttryckligen implementera och använda program för att skapa frågor som besvaras av en text: så kallad automatisk frågegenerering *(question generation QG)*. Frågegenerering är dock en systemtyp som kräver djupare analys än strängmatchning och statistiska lexikala modeller, vilka alltså ofta förekommer i system för informationssökning.

1.6 Språktekniska tillämpningar med djupare analysnivåer

Informationsökning är kanske den idag mest använda tillämpningstypen som kan kallas språkteknisk. Men det ska klargöras att det som i ett sökindex motsvarar ett skrivet dokument ofta representeras på ett rent kvantitativt sätt, som *en påse av ord* – en *bag-of-words* (term från Harris, 1954).[14] Det är ofta fråga om en relativt ”strukturlös” representation av ett dokument genom dess förekommande ord, och med data som t.ex. hur många gånger löpord finns med. Metoder med kvantitativa bag-of-words-modeller, når ofta resultat som är bra för uppgifter som indexbaserad *fritextsökning* och *automatisk textkategorisering*.

En annan analysmodell av textdata behövs för uppgifter som *grammatikkontroll* – eller framför allt i den prestigetyngda och från ett ekonomiskt perspektiv lockande uppgiften *automatisk översättning (maskinöversättning)*, vilken har benämnts den första icke-numeriska datamaskinella tillämpningen. Trots att denna tillämpningstyp kontinuerligt har utforskats sedan datorbegynnelsen och mycket stora ekonomiska belopp har satsats på forskning (speciellt ett sameuropeiskt vid namn *Eurotra*), är mänskliga översättare ännu inte riktigt hotade som översättare i fria domäner.

Uppenbart är i sådana automatiserade tillämpningar hur *positionen* för varje löpord i en text, till skillnad från i de tidigare nämnda modellerna har en potentiellt helt avgörande betydelse för resultatet.[15] Med andra ord krävs mer än en bag-of-words-representation av texten. Programmerad syntaktisk analys, *parsning*, i någon form, verkar ofrånkomligt. För ambitiösare AI-betonade

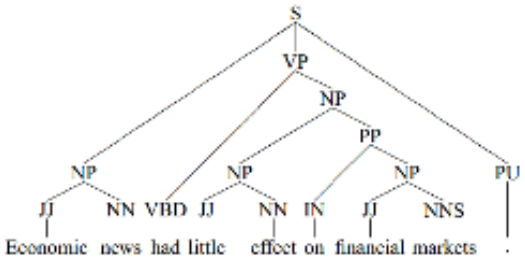
---

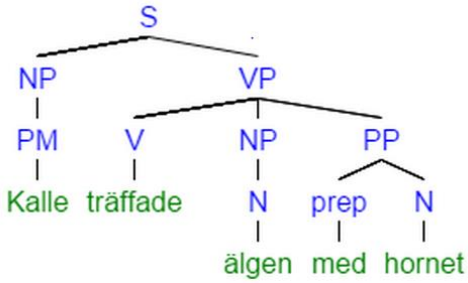
[14] “... for language is not merely a bag of words but a tool with particular properties which have been fashioned in the course of its use”, Harris (1954).

[15] Det går ju att rekonstruera olika möjliga texter av samma påse med ord, *Kalle lyssnar inte på kvinnan han känner / Kvinnan Kalle lyssnar på känner han inte,* osv. Dessa ”dokument” skulle kunna representeras lika i en bag-of-words-modell.

tillämpningar är analys på syntaxnivån (vilket oftast är resultatet av det som kallas *parsning*)[16] dessutom bara ett delsteg mot en mer semantiskt betonad analys och den ibland uttryckta målsättningen ”förståelse av texten.”

Enbart för svenska har försök till programmerad parsning av naturlig text förekommit sedan 1970-talet. Vad olika parsningsprogram som utvecklats för svensk text har åstadkommit i sina syntaktiska analyser är dock märkbart skiftande. Det finns grundläggande skillnader i själva de olika metoder som använts och de utdataformat för resultaten hos tongivande programmerade parsrar internationellt och för svensk text. Två typer av utdataformat exemplifieras nedan.





Figur 6 och 7: Två möjliga *frasstrukturella* parsningar (syntaktiska analyser) I analysen till höger: S: Sats, NP: nominalfras, PM: egennamn, VP: verbfras, V: (finit) verb, N: substantiv, prep: preposition. Detta är ett exempel på användning av en s.k. kontextfri grammatik (CFG).[17] Det blir här inte klargjort vad som är subjekt osv. Exemplet till vänster är hämtat från Nivre (2005).

```
<subjekt>Kalle</subjekt>
<pfv>träffade</pfv>
<objekt>älgen</objekt>
<adverbial>i affären</adverbial>
<tom>.</tom>
```

Exempel 1: En *funktionell* (dvs. med de grammatiska funktionerna subjekt, objekt etc.) parsning (huvudsatsanalys) genom *schemaparsning* (Wilhelmsson, 2010) genereras bl.a. i ett XML-format. *Pfv: primärt finit verb*

Åter till frågan: Vilka krav ställer då egentligen ett tänkt fungerande perfekt *frågebesvarande system (QG)?* Tillhör det den grupp av system som bör fokusera på lexikalt innehåll i likhet med många söksystem, eller innebär de försök till den djupare analys som kräver at texten parsas, som den i exemplen ovan? I detta arbete undersöks potentialen hos en modell som inbegriper den senare typen. – Är det motiverat att göra en (grammatisk) analys av texter för denna typ av

[16] Etymologiskt har termen *parsning* kommit från latin och ursprungets betydelse hänger samman med *dela*. En närliggande term i ett allmänt perspektiv som förmodligen klargör mycket för den som inte är bevandrad i området är att området helt enkelt behandlar *automatisk satslösning*. Till skillnad från i engelskan kommer de ”svenska” termerna parsning och parser här genomgående att användas för den datorimplementerade uppgiften, respektive programtypen. Det är också bara *syntaktisk* analys som avses här.

[17] Definitionen av ett kontextfritt språk (CFG) är formellt att det består av följande komponenter:
`Gcf = <N, T, S, P>`

`N`: Icke-terminaler
`T`: Terminaler (dvs. orden längst ner i trädet)
`S`: Startsymbol (S, dvs. Sats, alltså toppnoden i trädet)
`P`: Produktionsregler (Själva grammatiken, vilken består av omskrivningsregler med de ovanstående). En produktionsregel som används ovan är t.ex S → NP VP

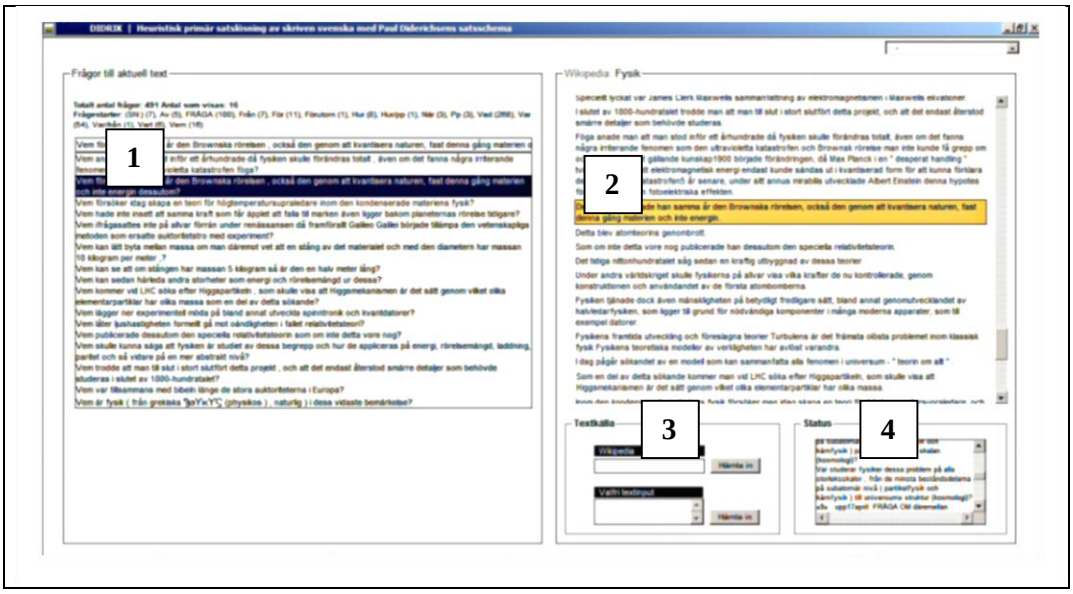
Generellt regelformat för produktionsregler: `A → ɣ`, där `A` är icke-terminal och `ɣ` är icke-terminaler eller terminal

informationssökning, och därigenom ge den ”strukturlösa” indata-texten en sorts struktur?

1.7 En motivering av frågegenerering för söksystem med naturligt språk

Resonemanget och figuren ovan har pekat på en viktig bakgrund till uppgiften frågegenerering (QG): Om de frågor en text kan sägas besvara kan ställas upp från början (t.ex. i en inläsningsprocess med denna förbehandling) av ett frågebesvarande system så behöver användaren idealiskt inte riskera att ställa frågor som ändå inte kan besvaras.

Vad som visats i ovanstående resonemang är hur ett fungerande system för QG i ett informationssökningssammanhang idealiskt vänder på uppgiften för ett frågebesvarande system på ett underlättande sätt: om det från början kan sägas vilka frågor som faktiskt kan besvaras av en text, per syntaktisk definition eller liknande (se kapitel 2), kan ett system ta fram besvarade frågor i explicit form från början. Detta är den bärande idén i frågegenerering (QG) för frågebesvarande system (QA). Målsättningen är då ett användargränssnitt *där enbart besvarade frågor kan ställas*. Den första prototypen för svenska som innebar att användaren enbart kunde ställa de frågor som genererade visades 2011 (Wilhelmsson, 2011).



Figur 8 Det grafiska gränssnittet i programmet består huvudsakligen av formulär för frågeval (t.v.) och texten som används (t.h.).

1. Inputfältet innehåller auto-komplettering för att snabbt välja en besvarad fråga
2. Texten som hela tiden visas för användaren, där svaret på en vald fråga scrollas fram och markeras
3. Val av artikel i *Wikipedia* eller annan godtycklig svensk textinput
4. Statusruta för diverse information under körning

I användning av detta program börjar användaren skriva och *väljer* alltså en fråga bland de som finns producerade i en auto-kompletteringsfunktion. När en fråga väljs scrollas den textmening som gav upphov till frågan (”svaret”) fram och markeras. Användningen innebär i ett GUI-perspektiv att användaren behöver skriva relativt lite eller inte alls. En generell målsättning är naturligtvis att användare så snabbt och enkelt ska finna svaret på ett visst informationsbehov. Här

finns förhoppningsvis i någon mening även de kvaliteter som Rogers et al. (2013, s. 215-217) tillskriver fungerande NUI-system, dvs. en naturlig eller intuitiv användning av systemet.

Var finner då denna programtyp sin plats i organisationen? Beynon-Davies (2013) gör en beskrivning av organisationers delar och tillhörande informations- och kommunikationssystem med olika roller. Stenmark (2002) behandlar speciellt *informationsåtkomst i intranät* i sin tidiga forskning. Detta är en tänkbar kontext för systemtypen i fråga, men det är inte heller nödvändigtvis så. (Ett tidigt namn på ett utvecklat system av Stenmark var för övrigt också just *Watson*, ej att förväxla med ovanstående QA-system.) Sökmekanismen med naturligt språk mot en databas av naturlig text skulle i fungerande skick kunna användas på många håll, och inte nödvändigtvis inom en organisation.

Oberoende av kontexten är det framför allt frågor om själva användandet av gränssnittstypen, dvs. *GUI/NUI*, enligt Rogers et al (2013) ovan, som ställs i detta arbete.

Prototypen ovan tjänar rollen som undersökningsobjekt i denna uppsats även om det inte används. Användarundersökningen som görs handlar om att ta fram autentiska frågor för texter i en användarstudie. Syftet, som beskrivs noggrannare i följande kapitel, är att utröna hur stor andel av sådana verkliga frågor som *skulle* kunna tas fram med aktuell metod och finnas med i systemet. För att beskriva vilka frågor som systemet kan, eller skulle kunna, generera tar följande kapitel upp hur aktuell frågegenerering för svenska fungerar. Detta är en beskrivning som kräver åtminstone en kort redogörelse för hur svensk grammatik ter sig i sammanhanget, för att förklara vilka frågor som kan genereras.

1.8 Syfte och frågeställning

Genom att uttryckligen generera frågor som den aktuella textmängden kan besvara bemöts idealiskt de två nämnda aspekterna hos söksystem med naturligt språk:

- Systemet fungerar inte längre genom att det söker efter svar på en sökfråga just när den ställs från en användare. Istället har en analys av texten, dess ’svar’ och motsvarande frågor redan gjorts på förhand.

- Eftersom frågor som besvaras av texten tagits fram är dessa frågor ramar för användningen. Användaren måste ställa en av just dessa frågor. Användarens okunskap om texternas innehåll, som diskuterades ovan, bemöts genom att andra frågor inte kan ställas i gränssnittet.

Betyder det att användningen av ett fungerande sådant system förbättrar utgångsläget för informationsinhämtning på ett idealiskt sätt?

- Syftet med denna uppsats är att utforska just relationen mellan *verkligt relevanta frågor* (enligt användare) och de som kan skapas (för närvarande, eller som skulle kunna utvecklas med aktuella metoder).

Det finns flera möjliga problem med det funktionssätt som skisserats som lösning på problemet ovan. Ett viktigt sådant problem är att frågegenereringen skapar ett visst frågeuttryck per ’svar’ som det finner i texten. Men varje fråga kan ställas på många olika sätt, med *varierande* ord- och syntaxval (se Beynon-Davies, s. 279). Är detta en hämmande aspekt i gränssnittet – betyder det att användaren så att säga ändå måste leta upp en existerande formulering av den fråga hon har i åtanke?

- Huvudfrågeställning: *I vilken grad kan de frågeformuleringar som riktiga användare väljer att ställa genereras automatiskt?*

När det gäller avgränsningar så är de inte helt tydliga från början eftersom studien delvis är explorativ. I detta arbete görs dock inte alltför djupa *språkliga* analyser av de insamlade resultaten.

1.9 Disposition för följande delar av uppsatsens

- *Kapitel 2* tar upp hur automatisk frågegenerering implementerats och den modell specifikt för svensk digital text som har nämnts. Detta kapitel beskriver de bakgrundskrav i fråga om språktekniska processer som implementationen är avhängig av. Detta kapitel tar även upp några frågeställningar om varför det är svårt att utvärdera ett sådant program, framför allt med en definierad målsättning som använts internationellt.

- *Kapitel 3* om studiens hypoteser presenterar vad som kallas en arbetshypotes för den empiriska undersökningen.

- *Kapitel 4, Metod,* är en beskrivning av metoden (empirisk undersökning) som valts för att besvara frågeställningen. Eftersom det är en studietyp som inte påträffats tidigare har den även en explorativ sida och sekundära frågeställningar.

- *Kapitel 5, Användarundersökning,* är en beskrivning av den användarstudie som företagits. Kapitlet tar upp de konkreta förutsättningarna för testpersonernas insatser och visar användarstudien i klartext.

- *Kapitel 6* redovisar resultatet från användarundersökningen. Det är en presentation av erhållen data utan långt dragna slutsatser.

- *Kapitel 7, Diskussion och slutsats,* tar upp resultaten i ljuset av hypoteser och inledande beskrivningar. Kapitlet innehåller några mer långtgående iakttagelser om studiens resultat och om förutsättningarna för frågegenerering i informationssökningsperspektiv.

Uppsatsen avslutas med referenser.

2 Automatisk frågegenerering (QG) som komponent i informationssökning

Användning av automatisk frågegenerering i ett informationssökningsperspektiv där användaren ställer frågor i naturligt språk bemöter som nämnts i föregående kapitel två omständigheter för frågebesvarande system. Det gäller för det första den intressanta relationen mellan frågor och besvarande segment och för det andra svårigheten (eventuellt omöjligheten) för ett system som inte gör djupare analys att veta om det över huvud taget har svaret på en viss fråga i sin textdatabas. Frågegenerering (QG) som en delkomponent i denna systemtyp försöker alltså bemöta svårigheterna på följande vis.

- I stället för att ett frågebesvarande system först under den egentliga användarinitierade sökprocessen *söker* möjliga svarsled för olika frågor, så *genereras* alltså besvarade frågor i klartext som ett allmänt föregående steg och är en typ av systematisk uppställning av befintlig information.

- I användning av frågebesvarande system, som de exemplifierade, vet egentligen varken användaren eller informationssystemet ifall svaret på en ställd fråga finns tillgängligt över huvud taget. Om systemet genom QG kan samla in besvarade frågor kan det begränsa användaren till att enbart ställa dessa besvarade frågor. Därmed skulle felaktiga svar eller enbart ”statistiskt bästa gissningar” som svar från systemet elimineras. Själva programmet får så att säga ett tentativt begrepp om vad det ”vet” respektive ”inte vet.”

Detta kapitel beskriver den prototypimplementation för frågegenerering från svensk text som idag existerar. Kapitlet inleds med en mycket kortfattad översikt av svensk satsgrammatik från ett funktionellt perspektiv, den s.k. positionsgrammatiken (Diderichsen, 1946). Anledningen till att detta behövs är att denna direkt ligger till grund för den metod för parsning (automatisk syntaxanalys) som används i den svenska implementationen, och dessutom för själva frågegenereringen, vilken i sin tur bygger på parsningen.

Beskrivningen av den tekniska processen (parsning och frågegenerering för fri text) leder fram till en viktig distinktion i detta arbete. Det är en karakterisering av en särskild mängd av frågeformuleringar per huvudsats i källtexten, här kallade *direktderivat* (se 2.3) som direkt skapas från textmeningarna (huvudsatserna). Dessa är i en mening de mest primitiva frågeuttrycken som kan genereras per text och per huvudsats (analysen och frågegenereringen sker satsvis).

Frågeuttrycken som samlas in i förprocessen kan sedan mångfaldigas genom omformuleringar och utökningar. Detta har idag bara delvis implementerats.

De mest primitiva frågeformuleringarna som skapas automatiskt för en text, *direktderivaten*, med pålagda formuleringsvariationer, och utökade med ytterligare, *härledda* frågor, har en huvudroll i detta arbete eftersom de utmejslar en gräns för vilken sammanlagd uppsättning av frågor som över huvud taget är möjliga att nå fram till (dvs. producera automatiskt) med den aktuella ansatsen av QG för svenska.

Det är denna frågeuttrycksmängd med frågeformuleringar per text som i kommande kapitel kommer att jämföras med verkliga frågeformuleringar som användare väljer att ställa i studiens användarundersökning. Om olika processer som adderar varianter av de skapade frågorna används blir det ofta en mycket stor samling av frågeuttryck. – Frågan är ändå: räcker samlingen för att fånga in verkliga användares naturliga frågor?

2.1 En minimal elementa i positionsgrammatisk syntax för svenska

Detta och nästkommande avsnitt innehåller en kort grammatisk beskrivning. Syftet med denna nödvändiga utflykt är att ge en redogörelse av hur frågegenerering för svenska behöver gå till.

Att göra en djupare analys (dvs. någon som urskiljer t.ex. en syntaktisk struktur och därmed åtminstone minimalt närmar sig motsvarande mänsklig analys) än den som enbart innebär varianter av strängmatchning tillsammans med olika statistiska modeller, måste utgå från språkets egen struktur. Hur är då den svenska språkstrukturen beskaffad? För att klargöra detta redogör följande avsnitt översiktligt för ett viktigt huvuddrag i svensk syntax, de positionsmässiga relationerna mellan de grammatiskt *funktionella* leden.

Svenska är liksom andra germanska språk, förutom engelska, ett V2-språk. Denna inte så vanliga egenskap innebär att en huvudsats finita (böjt i presens eller preteritum; t.ex. *talar/talade/ska/har*) verb kommer på plats 2 bland de funktionella led som bygger upp huvudsatsen. På satsinledande position 1 finns plats åt precis *ett* satsled (med vissa undantag, se Wilhelmsson, 2010, 3.6) – men detta led kan vara av olika slag.

	Satsbas (fundament)	**Finit verb**	Subjekt	(Sats-) adverbial	Icke-finit verb	Objekt, eg. subjekt, predikativ och objekt-liknande adverbial	Övrigt adverbial
		v	n	a	V	N	A
1:	*Kalle*	*skulle*	*[-]*	*nog*	*spela*	*en ny match*	*imorgon*
2:	*Imorgon*	*skulle*	*Kalle*	*nog*	*spela*	*en ny match*	*[-]*

Tabell 2. Satserna 1 och 2 i schemat visar två varianter (permutationer) av samma sats. Det som skiljer dem är att i 1 är subjektet *(Kalle)* i fundamentposition (första position), medan ett adverbial *(imorgon)* har placerats där i 2. Om ett led inte har flyttats fram först står det generellt på den position som anges i översta kolumnen. Bl.a. subjekt, objekt och adverbial flyttas fram på detta sätt och åstadkommer varianter av samma huvudsats *salva veritate* (med bibehållna sanningsvillkor, med i princip samma betydelse). Spåret ”[-]” markerar att ledet där ’flyttats fram’.

Det led som inleder huvudsatser är i deklarativ svensk text satsens subjekt i ungefär 60-80 % av fallen. De funktionella leden är grovt sett verb (finita, samt icke-finita verb), verbpartiklar, reflexiver (*sig* etc.), subjekt, objekt/predikativ och adverbial.

– I princip är detta grunden till en heltäckande beskrivning av de led som finns i omlopp i svenska satser. Tillsammans med positionsgrammatikens uppställningar av regler för de positioner som de olika ledtyperna kan inta, *syntaxen*, (nämligen i *satsschemat*, Diderichsen, 1946. m.fl.) ges en mycket kärnfull beskrivning av satsgrammatiken som nu är allmänt förekommande i grammatiska läromedel för svenska.

Inledare	Mittfält			Slutfält			
Satsbas (fundament)	Finit verb	Subjekt	(Sats-) adverbial	Icke-finit verb	Partikel-adverbial	Objekt, eg. subjekt, predikativ och objekt-liknande adverbial	Övrigt adverbial
Ni	*hade*	*[-]*	*nog*	*funnit*	*på*	*något nytt*	*nästa dag.*
Det man ville klargöra nu	*skulle*	*de som instämde*	*trots att de avsåg annat*	*tillskriva*		*egenskaper som fungerade*	*eftersom de inte kunde vänta*
Igår	*hade*	*det*	*faktiskt*	*passerat*		*en tankbil*	*på vägen.*

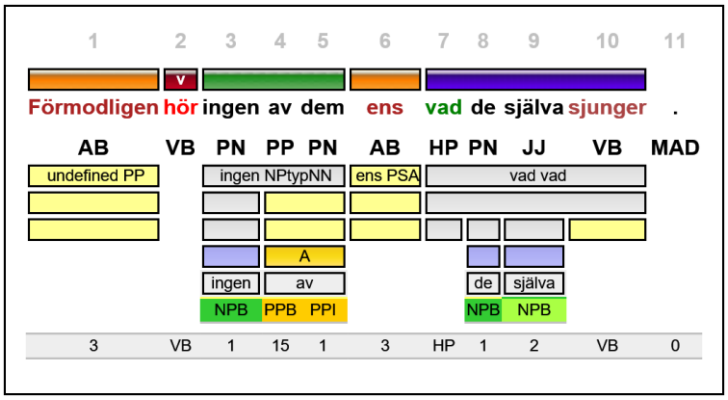
Tabell 3 *Svenska Akademiens språklära* (Hultman, 2003), sida 292, använder huvudsatsschemat (Diderichsen, 1946) som en grundbult i grammatikbeskrivningen. (Här delvis egna exempel.)

Det är lätt att missa poängerna med den visualisering som satsschemat innebär. Fundamentpositionen erbjuder alltså en möjlighet för de flesta led att placeras främst (spetsställning, fundamentering) och fångar in den naturliga variation som råder (*Ni hade nog funnit på något nytt nästa dag, nästa dag hade ni nog funnit på något nytt, något nytt hade ni nog funnit på nästa dag...)*

Den metod som används för parsning av fri svensk text, som första steg i frågegenereringen, är speciellt inriktad på att göra en sådan funktionell huvudsatsanalys. Parsningsmetoden kallas schemaparsning (Wilhelmsson, 2010). Utdataformatet från den exemplifieras nedan.

<subjekt>Ni som frågar om detta</subjekt> <pfv>hade</pfv> <adverbial>nog</adverbial> <adverbial>ändå</adverbial> <piv>kunnat</piv> <piv>köpa</piv> <objekt>en vän</objekt> <objekt>en present</objekt> <tom>.</tom>	<subjekt>Proportionell konsolidering</subjekt> <pfv>innebär</pfv> <objekt>att endast de egna andelarna i bolaget redovisas</objekt> <tom>.</tom>

Exempel 2: Utdataformatet i XML-format. Den huvudsatsanalys som den föreliggande metoden ger genereras bl.a. i ett XML-format. *Pfv: primärt finit verb* Exempel t.h.; hb09a-051, från *Stockholm Umeå Corpus* (Ejerhed, Källgren, & Brodda, 2006).

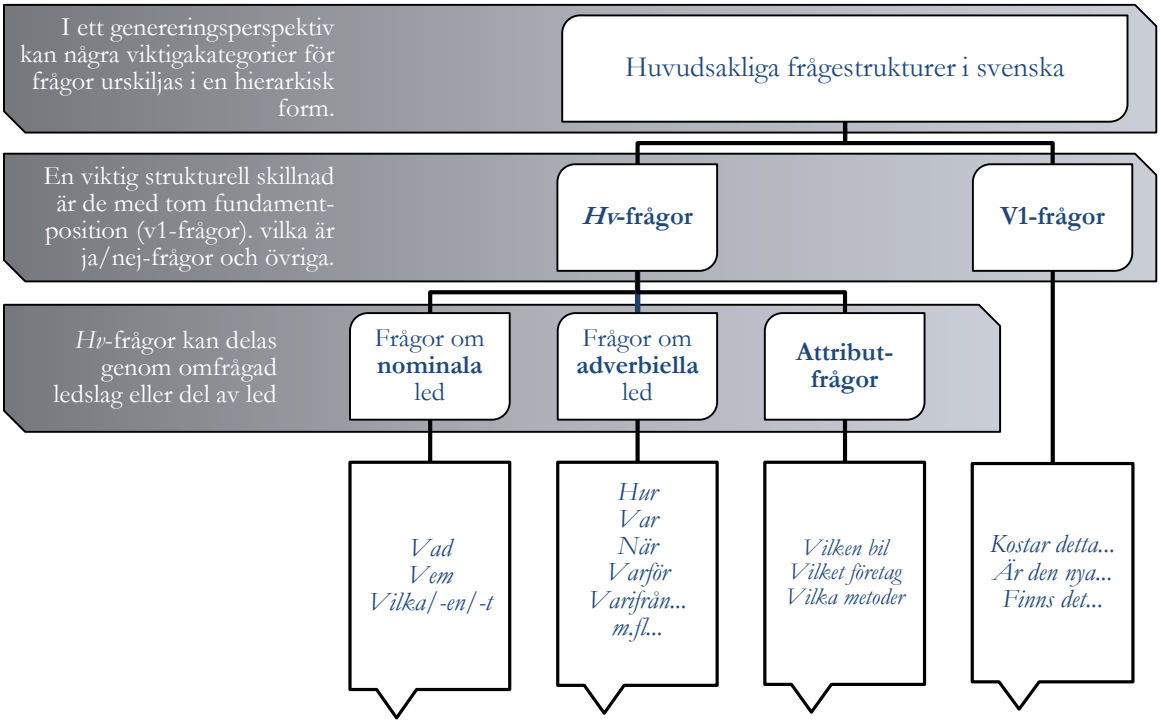


Figur 9 Analysen kan visualiseras i HTML (oförskönat exempel). Det är den övre raden ovanför texten som är det faktiska funktionella syntaktiska analysresultatet och som visas med färgkodning. Grönt: subjekt, blått: objekt, orange: adverbial, rött: verbled. Exempel: ja11-123 från *Stockholm Umeå Corpus (Ibid).*

Den analys som schemaparsningen ger (dvs. den struktur den kan sägas ge åt ”ostrukturerad data”) är alltså en grammatiskt funktionell analys på huvudsatsnivå.

2.2 Frågetyper

Frågor grupperas på varierande sätt i litteraturen. I genereringssammanhanget kan först en uppdelning mellan den formmässigt skilda V1-frågetypen och de andra göras. *V1*-frågor (verbinitiala frågor, med finit verb först, dvs. *ja/nej*-frågor) blir resultatet när att en deklarativ huvudsats får en tom inledande position, fundament (se Tabell 4). I frågegenereringssammanhanget kommer därefter den kvarvarande gruppen, s.k. *hv*-frågor att sedan delas upp på mer semantiska grunder.



Figur 10: *Hv*-frågor – namnet är ekvivalent med engelska *wh-questions* och innebär frågeord varav många tidigare har stavats med inledande *hv* i svenska. Bland *hv*-frågorna omfrågar *nominala* led subjekt och objekt/predikativ i en sats (oftast *vad/vem/vilket*). De *adverbiella* frågeleden är av ett betydligt större antal. Att avgöra rätt frågeord för ett godtyckligt adverbial är en icke-trivial uppgift (se Wilhelmsson, 2012). *Attributfrågor* gäller en *del* av ett led. Jämte dessa syns *V1*-frågor (*ja/nej*-frågor) t.h.

De former av frågor som strukturmässigt kan kategoriseras som *v1*-frågor eller *hv*-frågor är de som här främst kommer att beaktas, och som utgör majoriteten av frågeförekomsterna i skrift. Det är också dessa frågor som genereras av det aktuella QG-systemet för svenska. För redogörelsens skull bör nämnas att autentiska diskurser, speciellt i talspråk, innehåller frågor med fler, och mycket varierande (ofta förkortade) strukturella former, se Ericsson (2006).

2.3 Automatisk frågegenerering för svensk text: Produktion av direktderivaten

Hur sker då programmerad frågegenerering av det slag som diskuterats, och vilka är precis de frågor (frågeformuleringar) som skapas? Processen frågegenerering som här beskrivs skapar frågor per huvudsats i texten. Närmare bestämt skapas speciella frågeuttryck till ”svar” i texten. Även om en användare ställer en fråga som bevisligen besvaras av texten vill det ju faktiskt till att hon i sökgränssnittet (se Figur 8) hittar en formulering av just sin fråga bland de som har skapats.

Hur ska precis de uttryck som skapas hållas isär från formuleringsvarianter av samma fråga? För att kunna tala om de frågor – och mer precis – de *formuleringar* av frågor som är det direkta resultatet av den inledande relativt enkla transformationen direkt från ursprunglig sats till frågeform – används här den termen *direktderivat*. Ordet *fråga* har alltså en problematisk användning. Det är i naturligt språkbruk outrett om två närliggande formuleringar som har samma svar bör ses som samma *fråga*. Det är ju ofta *innehållssidan* (betydelsen) hos ett språkligt uttryck som är definierande för vad som är en och samma fråga. Således är det i vanligt språkbruk vanligt att betrakta ’*I vilka antal kommer tranorna till Hornborgarsjön varje år?’* och ’*Hur många flyttfåglar (tranor) var det som årligen brukar dyka upp vid Hornborgasjön?’* som en och samma

'*fråga'* fastän de har klart olika uttryck. Om bara ett av dem genereras kanske inte användaren av sökgränssnittet hittar det.

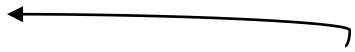
Olika frågeuttryck som är varianter av samma fråga (vilka har samma svar) måste här hållas isär för att klargöra att kanske bara ett genereras direkt och framträder i informationssökningen. Det är en avgörande distinktion att det ena av uttrycken skulle kunna vara ett direkt resultat av den grundläggande processen i frågegenerering, och därmed är ett *direktderivat* (och skulle alltså skapas av ett välfungerande QG-system), men inte andra uttryck (åtminstone initialt, utan vidare omformningsprocesser).

I nedanstående figurer och exempel redogörs för den grundläggande frågegenereringen, hur direktderivaten tas fram.

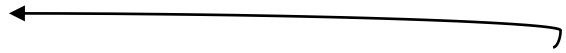
```
<subjekt>Uppföljaren</subjekt>
<pfv>motsvarade</pfv>
<adverbial>senare</adverbial>
<objekt>förväntningarna</objekt>
<adverbial>trots avbrotten</adverbial>
```

Exempel 3: en korrekt funktionellt analyserad (parsad) huvudsats i texten: följande steg visas nedan.

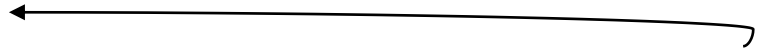
Uppföljaren *motsvarade senare förväntningarna trots avbrotten*



Senare *motsvarade uppföljaren [-] förväntningarna trots avbrotten*

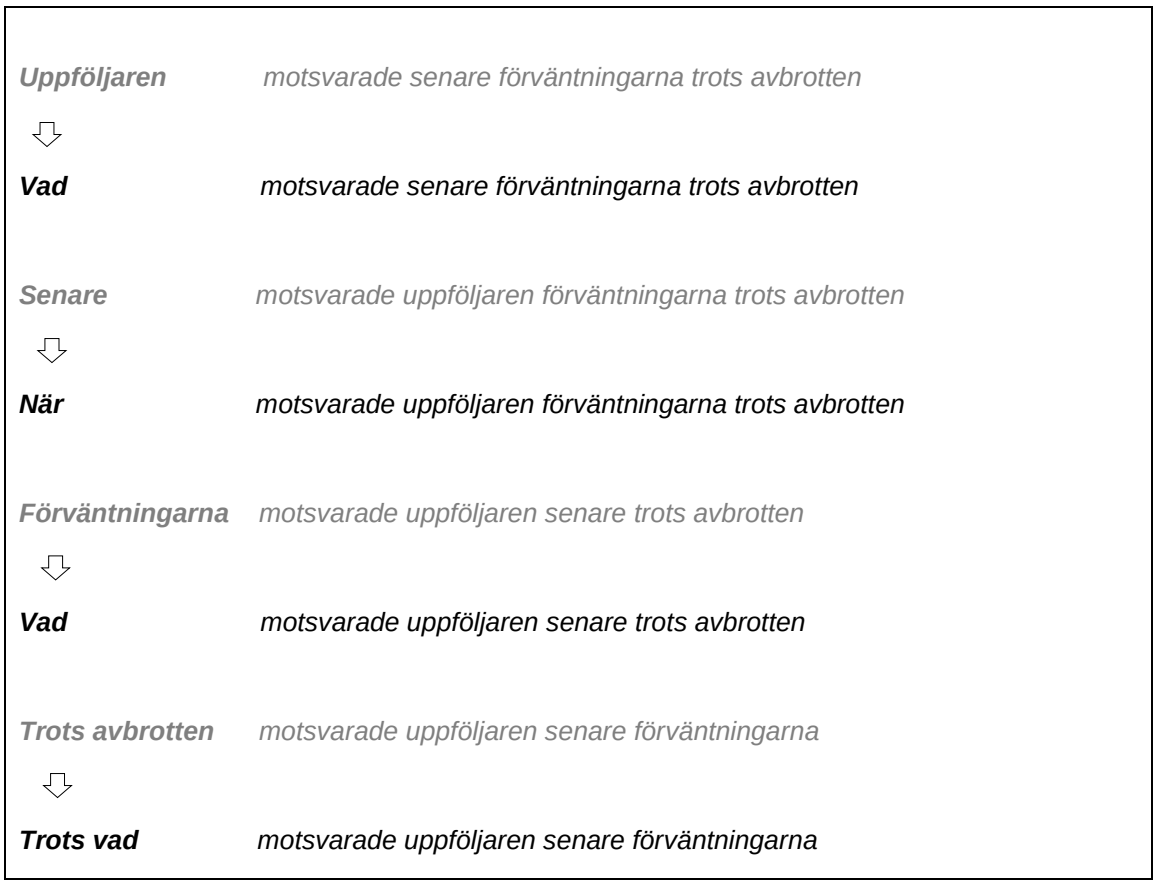


Förväntningarna *motsvarade uppföljaren senare [-] trots avbrotten*



Trots avbrotten *motsvarade uppföljaren senare förväntningarna [-]*

Figur 11: Insamlande av direktderivat (frågegenerering), steg 1. Ett första steg i processen efter parsningen är att placera varje flyttbart led enligt främst (fundamentering, spetsställning).



Figur 12: Insamling av direktderivat (frågegenerering), steg 2. Steg 2 innebär att byta ut det led som är framflyttat mot ett frågeord (*hv*-ord), eller frågesekvens som i nedersta exemplet.

Programmeringstekniskt är det inte speciellt svårt att skapa direktderivat (besvarade frågeuttryck) så som processen beskrivs ovan, förutsatt att parsningen är korrekt som i exemplet ovan. Det finns emellertid ett delsteg som har visat sig vara en felkälla: det sista steget då rätt frågesekvens, oftast *hv*-led ska väljas för att skapa en fråga (t.ex. *Hösten 1922 → när*), speciellt för adverbial.

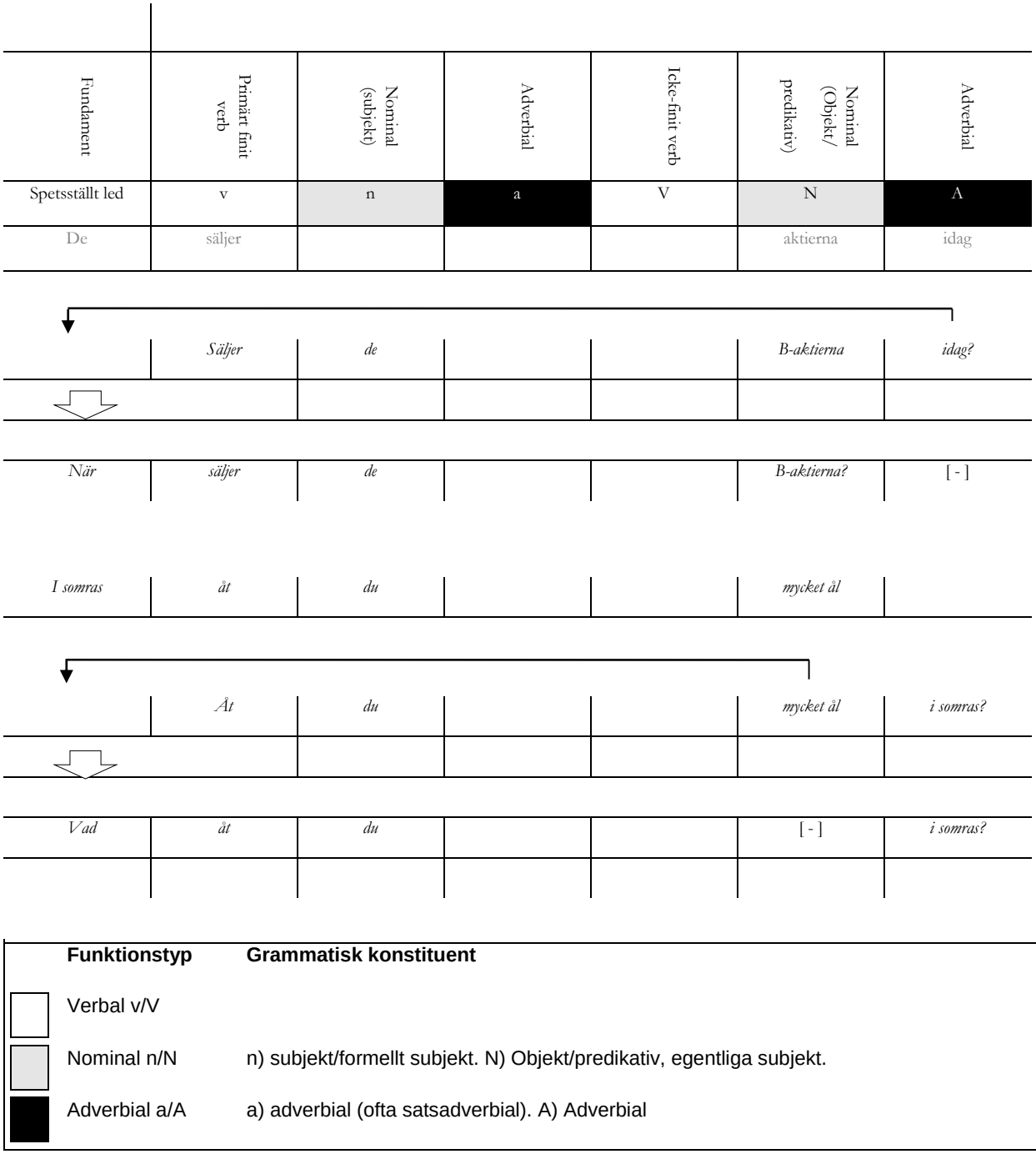
Det som här kallas *direktderivaten* är de frågeuttryck som skapas direkt utifrån en ursprungssats med ovanstående teknik. En sats som *Finansen* (subjekt) *påverkade* (v) *på börsen* (adverbial) ger följande direktderivat *Vad påverkade på börsen?* (subjektsfråga), *Var påverkade finansen?* (adverbialfråga). Detta sker alltså genom att det led som omfrågas flyttas till fundamentpositionen och därefter byts ut mot ett s.k. *hv*-ord.

Tre liknande frågekonstruktioner förutom de tydligaste ovan kommer också att kallas för direktderivat och kan eller skulle kunna genereras av ungefär samma maskineri.

1. I fallet *v1*-fråga (verbinitial fråga, *ja/nej*-fråga) görs första positionen (fundamentet) i stället tomt och leden står på sina andra positioner enligt satsschemat (se ovan). I Tabell 4 nedan exemplifieras V1-formen som ett delsteg. Genom att flytta det led som för tillfället finns på plats 1 till annan position, erhålls alltså en V1-fråga (*ja/nej*-fråga): *Finansen påverkade på börsen → Påverkade finansen på börsen?*

2. Det är i svenska även möjligt (men ovanligare) att skapa frågor utan att först flytta det omfrågade ledet till fundamentplatsen, dvs. bara genom att byta det omfrågade ledet mot ett *hv*-ord eller frågeordssekvens. Detta benämns *in situ*-fråga. *Finansen påverkade på börsen → Finansen påverkade var?*

3. Som direktderivat räknas här även *attributfrågor*. Attributfrågor innebär att en *del* av ett subjekt, objekt/predikativ eller objekt omfrågas. *De <u>nya</u> blå bilar som vi hade såldes → <u>Vilka</u> blå bilar som vi hade såldes?* (I huvudsatsobjektet, *de nya bilar som vi hade*, är det attributet (bestämningsordet) *nya* som här omfrågas).



	Funktionstyp	Grammatisk konstituent
	Verbal v/V	
	Nominal n/N	n) subjekt/formellt subjekt. N) Objekt/predikativ, egentliga subjekt.
	Adverbial a/A	a) adverbial (ofta satsadverbial). A) Adverbial

Tabell 4. Två ytterligare exempel på produktion av direktderivat, här visat i satsschemat. Denna gång är processen först att tydligt tömma fundamentet och visa motsvarande V1-fråga (*ja/nej*-fråga). Det översta innebär att *idag* omfrågas (ledet placeras i fundamentet, vilket först görs tomt, och byts mot *hv*-ord), i det nedre är det objektet *mycket ål* som omfrågas.

Det är alltså satser (huvudsatser) som är indata till QG-processen. Det finns möjlighet att ta fram fler huvudsatser än de explicita från samordnade finita verbfraser på huvudsatsnivå. Från *Syskonen*

kom hem och ställde in cykeln skapas två huvudsatser (*Syskonen kom hem* resp. *Syskonen ställde in cykeln*) – två olika satser har därmed producerats varifrån direktderivat kan samlas in.[18]

Frågeformuleringarna som är direktderivaten ”består” alltså till största del av de ord som förekommer i ursprungssatsen plus av frågeled som t.ex. *hv*-ordet *vad*. Denna första generation av genererade frågor kan efterföljas av andra frågetyper och frågeformuleringar som kan skapas med de förra som indata. Den nya samlingen frågor kan vidare ge upphov till många fler frågevarianter om dessa matas in i en process som skapar lexikal variation. Mängden frågeuttryck med formuleringsvariationer som skapas från en enda sats (dvs. de frågor inklusive formuleringsvariationer som satsen besvarar) kan bli förvånansvärt stor. Jämför Heilmann och Smith (2009) för ett engelskt system.

2.4 Hur många direktderivat ger en textmening upphov till?

Generering av direktderivat enligt proceduren ovan kan ge snabbt upphov till en mängd av frågor och närmast en explosion av formuleringsvarianter. Nedanstående textmening om Swaziland kan tjäna som exempel. Eftersom det är en samordning görs den först till två huvudsatser.

Landet blev 1894 ett protektorat under boerrepubliken Transvaal, ***och*** *kom under brittisk kontroll 1902*

1) Uppdelning: den samordnade finita verbfrasen i slutet ärver subjektet: två huvudsatser skapas

a) *Landet blev 1894 ett protektorat under boerrepubliken Transvaal* *b)* *Landet kom under brittisk kontroll 1902*

2) Reguljär generering av direktderivat från hela led

Vad (vilket land) blev 1894 ett protektorat under boerrepubliken Transvaal

När blev landet ett protektorat under boerrepubliken Transvaal

Vad blev republiken 1894

Vad (vilket land) kom under brittisk kontroll 1902

Under vilket lands kontroll kom landet 1902

När kom landet under brittisk kontroll

Exempel 4: I ovanstående exempel ger en textmening idealiskt upphov till sex besvarade frågor, direktderivat. Till dessa frågor kommer sedan två *V1*-frågor och i detta fall eventuellt någon attributfråga. När det gäller *in situ*-frågorna så skulle även ytterligare ca sex *formuleringar* kunna adderas.

En negativ aspekt med det blint mekaniska förfarandet ovan är att undermåliga frågor som inte normalt skulle ställas genereras (*Det regnade → Vad regnade* m.fl.). För den föreliggande studien har det ingen betydelse men det är troligt att en stor mängd ’dåliga’ frågor i ett användargränssnitt skulle kunna störa i ett frågebaserad informationssystem som använder frågegenerering. Heilman och Smith (2009) genererar ett överflöd av frågor, vilka de sedan försöker ranka automatiskt för att uppnå relevans.

[18] Det finns ännu fler sätt att ta fram satser för frågegenerering, men de kommer ej att behandlas ej här.

Går det då att säga hur många frågor och frågeformuleringar som en viss sats eller text kan ge upphov till med den mekaniska proceduren? När det gäller en enkel huvudsats är formeln för denna procedur generellt:

$$Antal\ direktderivat\ för\ en\ huvudsats = antalet\ omfrågningsbara\ led + 1\ (V1) + antalet\ attributfrågor$$

Formel 1: En grov approximation av det minsta antalet direktderivat per huvudsats.

Det är därmed ett antal som starkt varierar beroende på antal omfrågbara satsled i aktuell sats och hur många delar i satsled som är omfrågningsbara (attributfrågor).

När det sedan gäller omformuleringar (frågeuttryck) som varje direktderivat (fråga) kan ha så finns en risk för en fullständig explosion av antalet, vilket följande avsnitt illustrerar. Men frågan kvarstår: täcker de därmed in de frågor och uttryck som *riktiga* användare ställer och kallar relevanta?

2.5 Indirekta frågeformuleringar: Variationer på lexikal, syntaktisk nivå

Om ett QG-system ska kunna ge användaren viss valfrihet när hon formulerar sin fråga kan till att börja med rena ordmässiga utbyten beaktas. Att förändra texter genom synonymivariation har för svenska bl.a. undersökts av Rosell (2005) med hjälp av en *crowd-sourcing*-producerat synonymlexikon (Kann och Rosell, 2005). Det är rimligt att lexikala utbyten även skulle kunna vara användbart i frågor. Det är dock väldigt ont om helt utbytbara synonymipar när stilvärde beaktas. Dessutom är många ord polysema (t.ex. *bank* eller *fil*), vilket bl.a. har föranlett verksamhet inom forskningsfältet *word sense disambiguation* (*ordbetydelsedisambiguering* (Jurafsky & Martin, 2000). Resultat av lexikala utbyten för svensk frågegenerering har hittills inte varit speciellt uppmuntrande. Utan ordbetydelsedisambiguering görs lyckat utbyte i kanske bara hälften av fallen (se Wilhelmsson, 2010).

När det gäller syntaktiska (ordföljds- och konstruktionsmässiga) variationer så finns i svenska många möjligheter till nya uttrycksformer. Om transformeringarna utförs på grundsatserna i programmet kan naturligtvis frågeformuleringarna öka dramatiskt i antal. Det följande är exempel på generella variationer som befintliga svenska satser kan ha eller få.

Syntaktisk variation	Exempel
NA-rockad, främst lätta adverbial	*Han såg den inte* *Han såg inte den*
Konstruktion med formellt och egentligt subjekt	*En bil står parkerad på gatan* *Det står en bil parkerad på gatan*
Verb(-fras) i fundamentposition	*Han spelade fotboll* *Spelade fotboll gjorde han*
Passivisering/aktivering	*Mjölk dracks* *Man drack mjölk*
Utelämning av *ha/har/att/som*	*En man [som] vi känner*

Tabell 5: Några av de vanliga meningsbevarande transformationer finns bl.a. beskrivna i den forskning som följde den transformationsgrammatiska forskningen. Tabellen kommer närmast från Wilhelmsson (2010) modifierad. Typerna i tabellen där är i sin tur delvis hämtade från Holm och Larsson (1980) samt från Jörgensen och Svensson (1986)

Om huvudsatser i textmeningar transformeras enligt ovanstående regler mångfaldigas det antal kärnsatser varifrån direktderivat kan bildas. Som om det inte vore nog kan de ofta även kombineras. (De syntaktiska varianterna ovan är inte fullständigt implementerade hittills.)

En sista syntaktisk variation som bör nämnas i sammanhanget kallas *anaforlösning* (där speciellt pronomenresolution är en viktig del av forskningen) är ett mycket omforskat område både internationellt (t.ex. Mitkov, 1998) och svenska (Nilsson, 2010).

Kortfattat uttryckt handlar anaforlösning om att korrekt ersätta bakåtsyftade uttryck, ofta pronomen (*det, den, han, hon* m.fl.) med den korrekta referenten (*antecedenten*) i texten. *Han* i *Var reste han?*, som kan vara en genererad fråga, kan bytas ut mot den mycket mer användbara frågan *Var reste Amundsen?* (om *Amundsen* är den korrekta referenten). Resultaten från mångårig internationell forskning visar dock hur svår uppgiften är, både med regelstyrda ansatser (Fraurud, 1988) och med statistiska modeller (Mitkov, 1998).

De två följande avsnitten visar slutligen exempel på vad som räknas som utom räckhåll för den aktuella ansatsen. De metoder som följer är troligen möjliga att implementera, men de skiljer sig en del från de metoder för utökning av frågor och formuleringar som nämnts hittills. Skulle dessa typer av utökningar av frågeuttrycksmängden krävas för att täcka in de verkliga frågor som autentiska användare ställer?

2.6 Giltiga och ogiltiga logiska härledningar: Syllogismer, entymem och abduktion

Förutom det alldeles ordagranna textinnehållet och nämnda ’ytliga’ (i fråga om ordval och syntax) variationer är det rimligt att anta att mänskliga läsare drar slutsatser från texter på en djupare nivå.

Att det ska vara möjligt att dra logiska slutsatser utifrån texters innehåll, dvs. skapa andra, nya, satser förutom de som uttryckligen förekommer, ställer vissa krav på själva texten. Går det att skapa nya satser; logiska ”slut-satser” och använda även dessa för generering av fler frågor?

Det går att dra logiska programmeringsbara slutsatser från texter genom att förena innehållen i satser (*propositioner*) med varandra enligt logiska formler. På så sätt ”uppstår” t.ex. slutsats C från översats A tillsammans med undersats B nedan, fastän C inte fanns med uttryckligen i texten.

A Översats: *All persons born in the US are American citizens. […]*

B Undersats: *Barack Obama was born in the US*.

C Slutsats → *Barack Obama is an American citizen.*

Till skillnad från logiskt giltiga slutsatser enligt klassisk *syllogism*-form – som ovan exemplifierade regel (i detta fall *universell generalisering*) – förekommer i texters resonemang även s.k. *entymem*. Entymem skiljer sig från syllogismer då de strängt taget inte följer äkta logiska följdregler (Kjørup, 1996), Breitholtz (2010). Entymem definieras som en resonemangsform, som också skulle kunna sägas ha tre punkter, men där översatsen eller undersatsen utelämnats. Dessutom förlitar sig entymem i dagliga diskussioner ibland på t.ex. fördomar och människors kollektiva antaganden om olika företeelser.

Ett exempel på entymem såsom det förekommer i vanligt språkbruk skulle t.ex. kunna vara *”Det har regnat hela natten så nu får vi vara beredda med fiskespöna”.* Här finns en underförstådd, icke-uttalad, översats, kanske *”När det regnat på natten går fisken till”* – undersats och slutsats är det enda som egentligen sägs.

En annan egentligen logiskt ogiltig slutledningsteknik som förekommer inom vissa humanistiska discipliner är *abduktion* (se framför allt filosofen Peirce). Abduktion kan innebära en sorts associationsslutsatser eller ibland statistiskt sannolika slutsatser och har t.ex. tillskrivits romanfiguren *Sherlock Holmes* resonemangsstil. Utgångspunkten är att något faktum, någon händelse eller egenskap observeras – speciellt kan det vara fråga om något som överraskar. Detta faktum är något som kan beskrivas som ett resultat (en slutsats, dvs. C i exemplet ovanför). Abduktion innebär att ett A (en översats) *antas*, så att C därefter lätt kan förklaras. Exemplen på abduktion brukar tydliggöra att resonemangsformen känns tydligt ovetenskaplig:

> Alla mina bönor är vita. De kommer från säcken. →*(?)* Alla bönor i säcken är vita.
> (Peirce (1990) lätt justerat (Kjørup, 1996) s. 234)

Andra exempel som ges är *fåglarnas flykt tyder på regn* eller *handstilen visar skribentens slappa karaktär*. Detta visar naturligtvis på ett ganska svagt vetenskapligt bevisläge när abduktion tas till. Peirce har emellertid ett exempel som visar att regler som innebär abduktion finns ständigt närvarande: Dokument, och ting verkar peka på att den franske kejsaren Napoleon Bonaparte har existerat. Det godtas (av de flesta) som ett slags abduktivt bevis för att han faktiskt har funnits och är förklaringen till dessa spår.

Men vilka av de ovanstående härledningsteknikerna är praktiskt användbara för frågegenerering eller frågebesvarande informationssystem? I det översta härledningsexemplet i detta avsnitt, med en logiskt giltig deduktion blir resultatet gynnsamt för frågegenereringen. Den resulterande satsen om Barack Obama, C, är giltig och uppstår genom logisk härledning från texten, utan att den har förekommit explicit i texten. Det är tänkbart att programmatiskt skapa denna nya sats. Den kan alltså användas till att skapa fler frågor (och frågeformuleringar) som texten besvarar.

I fallen entymem och abduktion får det ses mer osäkert om någon av dessa metoders utvunna ’kunskap’ borde samlas in, t.ex. entymemets dolda över- eller undersats (såvitt känt finns detta inte heller implementerat i för fri text). Tillförlitligheten för ett sådant informationssystem (se 6.2) skulle inte säkert gagnas av detta.

2.7 Generering av fler frågor: lexikala/syntaktiska utökningar – ett mellanläge

Den ovanstående uppställningen av tänkbara utökningar av den automatiskt genererade mängden av frågor och frågeformuleringar har gått från det som här kallats *direktderivat* genom utökningar av formuleringar av frågor genom lexikala utbyten som synonymvariation och dessutom syntaktisk variation. Därefter har nu utvidgningar som skulle kräva en semantisk eller logisk analys exemplifierats.

Här ska nu avslutningsvis nämnas en viktig möjlig variation av frågeuttryck som i en mening tillhör de lexikala/syntaktiska variationerna, men som ses som ändå utom räckhåll beroende på speciella resurskrav.

En viktig form av syntaktisk/lexikala relationer är typen ”X *är författare till* Y” ↔ ”Y *skrev* X”. Dessa regler om samma information uttryckt på olika sätt inbegriper *både* en skillnad i ordval (lexikalt innehåll) och syntax (satsens form, i exemplet byter subjekt och objekt plats). Detta gör dem aningen komplicerade och svårfångade. Att dessa språkliga relationer ligger mellan lexikal och syntaktisk nivå påkallar behovet av en sådan gränsöverskridande resurs (troligen är ett s.k. *konstruktikon*, se Lyngfelt och Forsberg (2012), ett tänkt steg i denna riktning).

Ett frågebesvarande system av grundläggande slag kan därmed inte koppla samman *Vem är författare till Bilbo?* med det potentiella svarssegmentet *Bilbo skrevs av Tolkien*. Lin och Pantel (2001) tillhör de som har utforskat denna möjlighet till utökning.[19] Dessa utökningar/transformationer kan sannolikt vara mycket användbara för informationssystem med frågor i naturligt språk. De satspar som Lin och Pantel för samman är ibland ekvivalenta, och ibland härledningar som är *(icke-symmetriska)* relationer, dvs. de som bara kan göras åt bara det ena hållet *(far till → släkt med)* – se 7.1.

2.8 Utvärdering av implementationer av frågegenerering

Inom det internationella fältet QG (där engelska ofta har en särskild roll) har det funnits en något oklar definition rörande systemtypens målsättning i omlopp. Uppgiften som QG ska lösa har ibland beskrivits som ’att skapa samtliga frågor som en text’ besvarar. Med tanke på att det måste anses omöjligt att göra en uttömmande sådan uppställning (speciellt när det beaktas att olika användare besitter olika kunskaper och därmed kan dra olika nya slutsatser med hjälp av texten).

Om det inte går att fastställa hur många frågor det finns totalt och vilka de är, så innebär det också att *andelen* genererade frågor av det totala antalet är okänt. I konferenshandlingar editerade av Rus & Graesser (2009) används en definition för själva genereringsuppgiften enligt nedanstående formel. I uppgiften ingår med andra ord att koppla fråga till textsegment.

[19] De inledande resultaten där visar på svårigheten i uppgiften.

Given:

- Text $\boldsymbol{T}$

Create:

- Text-Questions pairs $\{P_1 \ldots P_n\}$ each represented as a (K_i, Q_i) pair, where K_i, the target text, indicates which text segment from $\boldsymbol{T}$ represents the answer and the Q_i represents a question that would elicit K_i

Formel 2: En definition av själva fråga-till-text-uppgiften i som ett steg i genereringsuppgiften, i handlingar editerade av Rus & Graesser, 2009)

Användarundersökningen här är en utvärdering i fråga om potentiell kapacitet för den modell som finns. Utvärderingen gäller alltså systemets tänkta tillräcklighet i den konkreta användarsituationen.

3 Hypoteser rörande användarstudien

Vetenskapen, speciellt naturvetenskapen, söker i empirin lagbundenheter. Metoden att gå från hypotes till en sådan allmän regel brukar ofta företas genom iakttagelser av enskilda fall, t.ex. genom experiment, tillsammans med *induktion* (dvs. att gå från enskilda fall till uttalanden om hela mängden): den *hypotetisk-induktiva* metoden. Induktion är till skillnad från deduktion (dvs. att använda en regel, att gå från kunskap om alla till att uttala sig om den enskilda instansen) egentligen inte en giltig logisk manöver. Induktion gäller alltså istället språnget från enskilda fall (t.ex. iakttagelser av en viss fågelart, t.ex. svanar) till allmänna uttalanden om t.ex. *alla svanar*. Det är teoretiskt sett inte alls oproblematiskt hur detta steg i kunskapsgenerering sker eller borde ske.

Allmän vetenskaplig lag	Hypotes (förslag till allmän vetenskaplig lag)
↓ *Deduktion*	↑ *Induktion*
Enskilda fall	Enskilda fall

Popper (1935) försöker klargöra att *induktionen*, vägen från de enskilda fallen till den allmänna regeln, nog inte ska ses som huvudsysselsättningen hos empirisk vetenskap: [Hypoteser om de universella lagbundenheterna] ”kan bara nås genom intuition, baserad på en form av inlevelse i erfarenhetens objekt”, (Kjørup, 1996), s. 86). Popper menade, inspirerad av Albert Einstein, att hypoteser rörande den allmänna regeln får tas fram med en så ’ovetenskaplig’ metod som intuition – men att denna regel sedan testas genom att göra förutsägelser om enskilda fall och kontrollera så att de stämmer med vad regeln säger. Denna efterföljande prövning av regeln/hypotesen, vilken kontrolleras, är alltså av äkta logisk deduktiv karaktär. När den tentativa regelns förutsägelser prövas mot utfall i verkligheten, t.ex. i experiment, stjälps den om dess förutsägelser visar sig felaktiga: hypotesen visar sig otillräcklig. För varje korrekt förutsägelse stärks den istället.

3.1 Primär frågeställning och en arbetshypotes

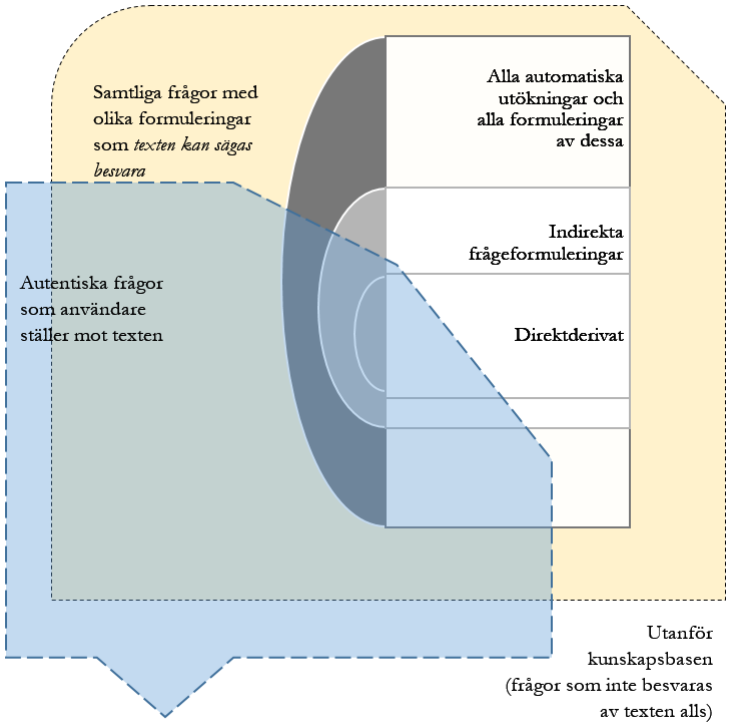
I detta arbete får möjligheten ’att de frågor som autentiska textläsare väljer att ställa är sådana att de skulle kunna skapas mekaniskt (närmare bestämt genom beskrivna utökningar av den aktuella metoden för svenska)’ spela rollen av en sådan induktivt framtagen hypotes. Det bör sägas att det inte finns något krav på att hypotesmakaren verkligen *tror* att den är en sådan allmän lag. Det kan rättare benämnas *arbetshypotes* (Patel och Davidson, 1994, s. 19).

Undersökningens huvudfråga gäller alltså – tillspetsat uttryckt – om *samtliga* frågor från användarundersökningen skulle kunna skapas mekaniskt på detta sätt. Det behövs endast en enda användarformulerad fråga som inte är sådan för att motbevisa hypotesen. För att *bevisa* att en hypotes verkligen gäller som en allmän lag och *verifiera* den, enligt Popper, måste hypotesen kunna prövas i upprepade experiment.[20]

Det egentliga *syftet* med den aktuella undersökningen är emellertid vidare än de helt konkreta resultaten i form av arbetshypotesens styrkande eller vederläggning. För praktiska ändamål, som

[20] Men hur många sådana experiment krävs – när kan hypotesen ses som en lag? Om uttalandet verkligen gäller *alla* fall, så når upprepade empiriska försök aldrig fram till en riktig teoretisk *verifiering*, menade Popper (Kjørup, 1996), s. 86.

frågan om användbarhet av metoden frågegenerering i praktiska informationssystem som här, är frågeställningen ändå mycket användbar.



Figur 13: Frågeställningen gäller de autentiska informationssökande frågornas art. I vilken utsträckning ligger de autentiska frågorna från användarundersökningen inom det genereringsbaras ramar?

Det är i själva verket minst lika intressant att veta i ett tekniskt utvecklingsperspektiv *hur stor andel* av samtliga användares frågor som de eventuella undantagen utgör, grovt sett. Det är vidare mycket intressant att studera hur undantagen ser ut. Hur kommer det sig, om det är fallet, att användare i så fall inte ställer frågor som tillhör direktderivaten med utökningar? (Här finns alltså några bakomliggande forskningsfrågor som ligger utanför denna undersöknings egentliga gränslinjer: hur går det till när riktigare användare ställer frågor, och varför kan somliga eventuellt inte genereras helt mekaniskt?)

3.2 Läs inte mellan raderna – det står inget där

Uppenbart är hur de blint producerade direktderivaten med utökningar dels är ”korrekta” genom sin systematiska tillkomst, men också kan vara en rejält naiv uppfattning om det totala informationsinnehållet hos texten. Innehåller texten inte någonting mer än detta – är de automatgenererade frågorna verkligen (något som ens liknar) en uttömmande informationsredogörelse? Men den som ställer denna fråga måste i så fall peka på vilka andra frågor (och formuleringar) än dessa som texterna då besvarar.

Det aktuella arbetet är ett direkt inlägg i detta spörsmål. Syftet med undersökningen kan på ett sätt ses just som att lyfta fram dessa *andra* frågor och frågeformuleringar – och vidare de som anses vara relevanta och som faktiskt uttrycks av mänskliga läsare. Är de autentiska frågorna kanske helt skilda från de mekaniskt producerade, och hur ser de i så fall ut? Om de uppstår från något ”mellan raderna” – hur besvaras de då och vad skiljer dem från de som är inom räckhåll för att skapas automatiskt?

3.3 Sekundära frågeställningar: autentiska frågors kvantitativa fördelning och 'informationsrika' delar av texten

Med tanke på att denna systemtyp för svenska aldrig hade utvärderats förut uppkom många andra frågeställningar förutom den om de skapade frågornas genererbarhet. En mycket intressant undersökningsfråga gäller hur naturliga frågor fördelar sig syntaktiskt. Går det att urskilja mönster när det gäller vilka frågeslag som används i ett autentiskt sammanhang? Denna sekundära frågeställning är (såvitt känt) outforskad.

Ett annat spörsmål i denna undersökning är om det verkar finnas vissa segment i texten som är speciellt omfrågade (dvs. som innehåller särskilt många svar på frågor). En liknande fråga är om undersökningsdeltagarnas frågor sammanfaller. Är det möjligt att förutsäga vilka segment som anses vara relevanta?

Avslutningsvis ska en fråga om uppgiftsformuleringen för användarna i undersökningen som beskrivs i nästa kapitel tas upp. I detta arbete har flera rent terminologiska spörsmål uppkommit. T.ex. har det blivit nödvändigt att mynta en term som *direktderivat* (se 2.3) för att kunna skilja mellan likartade frågeformuleringar som i dagligt språkbruk ändå skulle riskera att kallas för 'samma fråga.' – Typisk gäller det frågeformuleringar som har samma svar men som ändå kan ha uttryckts mycket olika; *När stänger den stora livsmedelsbutiken på tisdag/Hur sen är öppettiden för Coop imorgon?*

Det ska dock klargöras att detta begrepp problematiserats mycket inom informationsvetenskapen och inom lingvistiken, närmare bestämt inom pragmatiken (Grice, 1975).

4 Metod

För att utreda uppsatsens frågeställningar rörande autentiska och konstgjorda frågor är den metod som valts en webbaserad undersökning. Syftet är att skapa en kvantitativ bild av vilka frågor och frågeuttryck som autentiska användare väljer att ställa gentemot tre texter och som de dessutom anger som relevanta frågor. Undersökningsdeltagarna ombeds även att ange de textmening(-ar) i texterna, som besvarar de frågor som de har valt att ställa.

Eftersom undersökningsdeltagarna markerar upphovet (svaret) till sina frågor i texterna underlättas efterarbetet såtillvida att det blir klart huruvida ett angivet frågeuttryck tillhör den klass av uttryck som här kallas direktderivat eller inte.

Den aktuella undersökningen är den första i sitt slag och är främst en kvantitativ mätning, men den har även en explorativ sida och syftar egentligen till mer än att bara bevisa (genom genereringsbara frågor bland de autentiska) eller motbevisa (genom icke genereringsbara frågor bland de autentiska) arbetshypotesen. Patel och Davidson (1994) beskriver en explorativ undersökning som en som rör ett område med kunskapsluckor, där målet är så mycket kunskapsinsamling som möjligt (s. 11).

4.1 Användarundersökningens textmaterial

I fråga om urval av texter har dessa valts ut för uppfylla kraven på informationsrikedom som ett tänkt fungerande system behöver. Innehållet i texterna är faktamässigt och de beskriver inte tidsförlopp där olika försanthållanden gäller i olika stycken. Styckena är vidare av relativt kort längd för att deltagandet i studien inte ska behöva ta för lång tid. I likhet med flera nämnda studier av svensk frågegenerering och internationella ansatser för informationsextraktion används texter från *Wikipedia*.

De tre texterna som använts i användarundersökningen kommer från fritt tillgängliga Wikipedia-artiklar som behandlar *a) en plats, b) en person* och *c) en mer abstrakt företeelse*. Två av texterna behandlar ämnen som troligen är välkända för läsarna medan det tredje ämnet förmodligen är okänt för undersökningsdeltagarna. Syftet med ämnesspridningen är att om möjligt stävja eventuell informationsmässig snedvridning som skulle kunna råda hos någon ämnestyp. Texterna redovisas i klartext nedan.

4.2 Finns svårigheter med den aktuella metoden?

Idealiskt skulle deltagarna i undersökningen förmodligen vara lika väl representerade i fråga om ålder och kön. Detta var dock inte möjligt att styra i ett utskick om frivilligt deltagande, där alla svar har betraktats som värdefulla och intressanta. För undersökningens huvudfråga är den avgörande kvaliteten att frågorna, oavsett upphovsman är *autentiska frågor*, som kan jämföras med de automatgenererade.

Användarstudiens upplägg har inneburit att instruktionerna har varit tämligen korta (se nedan) och inte alls för deltagarna avslöjar de bakomliggande syftet med uppgiften att producera relevanta frågor som texterna besvarar.

Studiedeltagarna ska ’producera vad som uppfattas som relevanta besvarade frågor (frågeformuleringar)’ utifrån en text. Undersökningens primära syfte är att se i vilken grad de

autentiska frågeformuleringarna befinner sig inom nuvarande frågegenereringssystems råmärken. Det kan självklart vara så att texterna eller undersökningsdeltagarna på något sätt innebär en snedvridning *(bias).* Hur skulle detta kunna undvikas?

När det gäller undersökningsdeltagarna så har det alltså inte varit möjligt att kontrollera vilka som genomfört undersökningen, då studien är anonym. Köns- och åldersfördelning har dock samlats in och redovisas i 5.2. När det gäller texterna och deras representativitet finns här en känd problematik som förekommer när språktekniska system ska testas: vad är egentligen en bra representation av ’allmän svensk text’? Hur är den kvantitativt fördelad över genre etc.? När frågan gäller publicerad svensk text finns en känd textkorpus, *Stockholm Umeå Corpus* (Ejerhed, Källgren, & Brodda, 2006), som har sammansatts med syftet att representera de genrer av publicerad 1990-talstext som en svensk person konsumerar. Det skulle nog vara svårt att använda denna textkorpus av flera skäl. Den är uppdelad i nio genrer med undergenrer. Det skulle troligen göra textvalet svårt med tanke på uppgiften. Somlig text är skönlitteratur, vilket lämpligen undviks i detta sammanhang. Faktatext från *Wikipedia* kan anses mer realistisk i rollen som textdatabas.

5 Användarundersökning

Användarundersökningen genomfördes i form av en webbaserad enkät (se nedan) med anonyma undersökningsdeltagare. Uppgiften har varit att läsa tre korta texter och ställa relevanta frågor utifrån dem (se de precisa formuleringarna nedan). Den enskilde undersökningsdeltagaren har ombetts att formulera tre frågor per text (sammanlagt nio frågor per person). Instruktionen innebar även att för varje fråga som formulerades ange var i källtexten (textmeningsnummer) som svaret på frågan kunde erhållas. (Syftet var som nämnts att därigenom snabbt kunna sortera frågorna mot de textsegment (svar) de korresponderar mot för att underlätta resultatanalysen.)

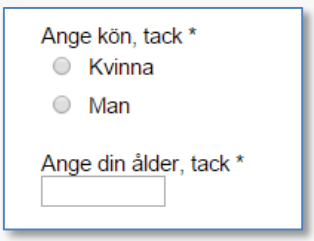
5.1 Användarundersökningen i klartext

Den Internetbaserade undersökningen genomfördes med webb-tjänsten *QuestionPro Survey*. Det nedanstående är en redovisning av undersökningen i klartext. De tre textsegmenten från Wikipedia som valdes ut handlar om platsen *Visby*, personen *Julio Baghy* respektive begreppet *gymnasieskola*. Patel och Davidson (1991) har följande att säga om de medföljande instruktionerna.

> ”Om vi använder oss av enkäter, finns bara ett sätt att motivera individerna och det är genom det brev som medföljer enkäten, det s k *missivet.* Detta missiv måste innehålla all information som vi vill ge och det är alltså mycket viktigt att det är korrekt utformat.” (Patel och Davidson, 1991)

I anvisningen till den aktuella användarundersökningen förekommer begreppet *relevans*. Att ett svar/fråga/dokument är *relevant* (utan vidare specificering) exemplifierar en modern språkanvändning som används oproblematiserat i detta sammanhang. Användarna ombeds att framföra ’relevanta’ frågor om texterna – och syftet är att därmed få användbara, viktiga, realistiska frågor, frågor som gärna får ses som prominenta i de olika textinnehållen. Denna formulering av instruktion övervägdes noga före undersökningen. Den användes utan att problematiseras.

Välkommen till denna anonyma undersökning som rör kunskap och informationsinnehåll i faktatext, ett område som undersöks vid Göteborgs universitet. Jag är mycket tacksam om du fullföljer hela webbenkäten, avbryt helst inte. Påbörja enkäten när du vet att du kan ägna ca 15 min åt den. Undersökningen går ut på att läsa 3 korta texter och skriva ner frågor. Det ska vara frågor som besvaras av texten och som du upplever är relevanta. Efter varje fråga anger du precis var i texten som svaret på din fråga finns (meningens nummer). Sprid gärna denna undersökning!



Innan du startar kan du se på ett exempel på ungefär hur ett svar kan se ut: Exempel

1 Ipad (i kommersiella sammanhang iPad) är en surfplatta, eller en portabel pekdator, från det amerikanska hemelektronikföretaget Apple. 2 Ipad presenterades första gången den 27 januari 2010 av Apples grundare och VD Steve Jobs på en presskonferens i San Francisco. 3 Ett stort hemlighetsmakeri hade föregått presentationen, tidiga rykten benämnde produkten Islate eller Apple Tablet. 4 På den amerikanska marknaden var det initiala priset för en Ipad mellan 499 och 829 dollar beroende på minnesstorlek och 3G-uppkoppling och på den svenska marknaden 4 695 kronor inklusive moms för den billigaste.

Källa: Wikipedia

Fråga 1 - Skriv en relevant fråga som texten besvarar.

När presenterades IPad för första gången?

Skriv ner den mening som besvarar fråga. Om det inte går att använda bara en mening, skriv första meningen följt av ett f

2

Fråga 2 - Skriv en relevant fråga som texten besvarar.

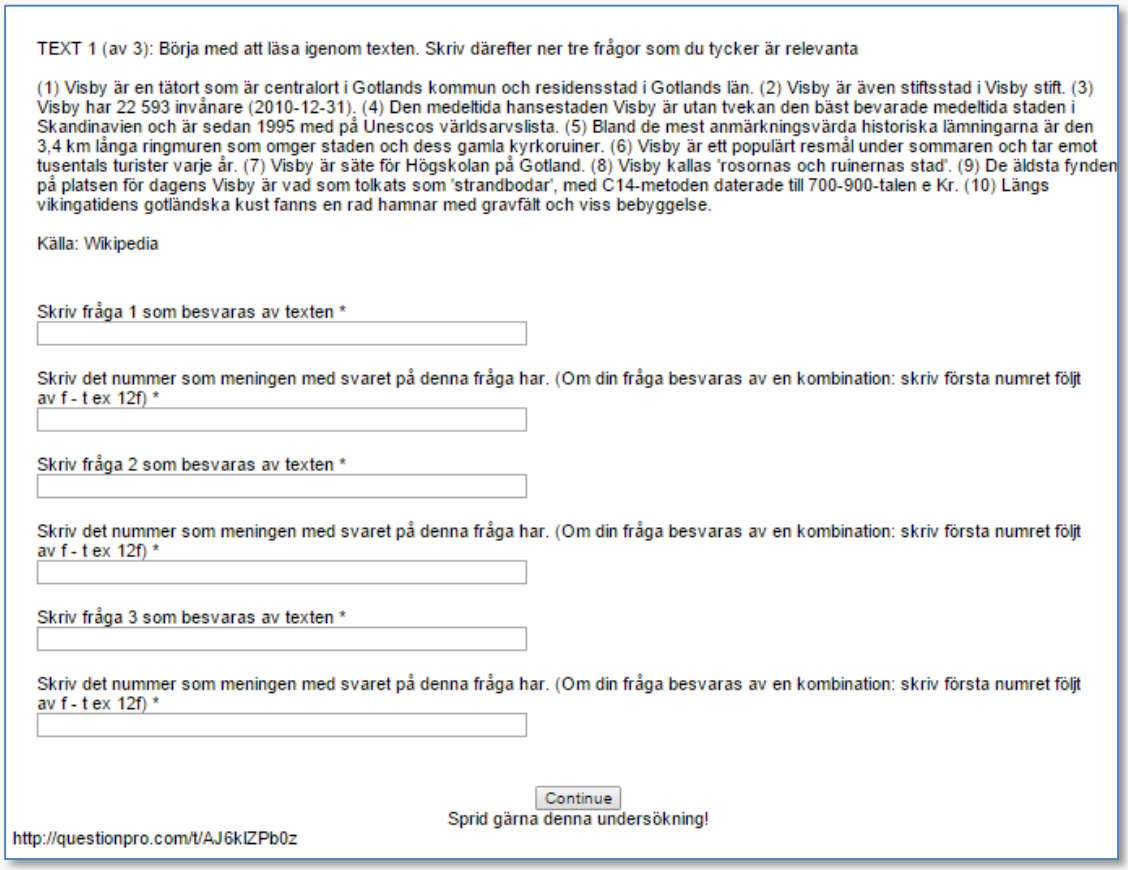
Vad kostade Ipad från början?

Skriv ner den mening som besvarar fråga. Om det inte går att använda bara en mening, skriv första meningen följt av ett f

4

Ovan: Exempeltext (från artikeln iPad) i instruktionen i webbundersökningen

Nedan: Den första av de tre texterna till undersökningsdeltagarna



TEXT 1 (av 3): Börja med att läsa igenom texten. Skriv därefter ner tre frågor som du tycker är relevanta

(1) Visby är en tätort som är centralort i Gotlands kommun och residensstad i Gotlands län. (2) Visby är även stiftsstad i Visby stift. (3) Visby har 22 593 invånare (2010-12-31). (4) Den medeltida hansestaden Visby är utan tvekan den bäst bevarade medeltida staden i Skandinavien och är sedan 1995 med på Unescos världsarvslista. (5) Bland de mest anmärkningsvärda historiska lämningarna är den 3,4 km långa ringmuren som omger staden och dess gamla kyrkoruiner. (6) Visby är ett populärt resmål under sommaren och tar emot tusentals turister varje år. (7) Visby är säte för Högskolan på Gotland. (8) Visby kallas 'rosornas och ruinernas stad'. (9) De äldsta fynden på platsen för dagens Visby är vad som tolkats som 'strandbodar', med C14-metoden daterade till 700-900-talen e Kr. (10) Längs vikingatidens gotländska kust fanns en rad hamnar med gravfält och viss bebyggelse.

Källa: Wikipedia

Skriv fråga 1 som besvaras av texten *

Skriv det nummer som meningen med svaret på denna fråga har. (Om din fråga besvaras av en kombination: skriv första numret följt av f - t ex 12f) *

Skriv fråga 2 som besvaras av texten *

Skriv det nummer som meningen med svaret på denna fråga har. (Om din fråga besvaras av en kombination: skriv första numret följt av f - t ex 12f) *

Skriv fråga 3 som besvaras av texten *

Skriv det nummer som meningen med svaret på denna fråga har. (Om din fråga besvaras av en kombination: skriv första numret följt av f - t ex 12f) *

Continue

Sprid gärna denna undersökning!

http://questionpro.com/t/AJ6klZPb0z

Nedan visas de tre texterna tydligare i en och samma tabell (i undersökningen visades texterna alltså en åt gången per sida, enligt ovan).

Text 1, Visby	*Text 2, Julio Baghy*	*Text 3, Gymnasieskola*
(1) Visby är en tätort som är centralort i Gotlands kommun och residensstad i Gotlands län. (2) Visby är även stiftsstad i Visby stift. (3) Visby har 22 593 invånare (2010-12-31). (4) Den medeltida hansestaden Visby är utan tvekan den bäst bevarade medeltida staden i Skandinavien och är sedan 1995 med på Unescos världsarvslista. (5) Bland de mest anmärkningsvärda historiska lämningarna är den 3,4 km långa ringmuren som omger staden och dess gamla kyrkoruiner. (6) Visby är ett populärt resmål under sommaren och tar emot tusentals turister varje år. (7) Visby är säte för Högskolan på Gotland. (8) Visby kallas 'rosornas och ruinernas stad'. (9) De äldsta fynden på platsen för dagens Visby är vad som tolkats som 'strandbodar', med C14-metoden daterade till 700-900-talen e Kr. (10) Längs vikingatidens gotländska kust fanns en rad hamnar med gravfält och viss bebyggelse.	(1) Julio Baghy (ungerska Baghy Gyula), född 13 januari 1891 i Szeged, död 18 mars 1967, var en ungersk författare och skådespelare. (2) Han lärde sig esperanto 1911 och började arbeta för esperantorörelsen under en sexårig krigsfångenskap i Sibirien. (3) Han var en framstående esperantoaktivist och lärare, och under en tid vice ordförande för Esperantoakademin. (4) Baghys far var skådespelare och hans mor teatersufflör. (5) Efter skolgången blev även han skådespelare och regissör vid olika teatrar. (6) Kriget avbröt hans karriär och han tvingades tillbringa sex år utanför sitt hemland i rysk krigsfångenskap. (7) Redan under hans ungdom publicerades många av hans dikter och noveller i ungerska tidskrifter. (8) 1911 lärde han känna esperanto, och dess inre idé (interna ideo) lockade honom omedelbart. (9) Hans omfattande engagemang för esperanto började redan i fånglägret i Sibirien, där han ledde flera kurser för människor från olika länder. (10) Sedan han återvänt till Ungern efter kriget blev han en av de viktigaste ledarna för esperantorörelsen: han höll många kurser på olika nivåer, ledde Esperanto-Rondo Amika, arrangerade litterära kvällar och så vidare.	(1) En gymnasieskola är en avgiftsfri och frivillig sekundärutbildning i Sverige för ungdomar som har gått ut grundskolan. (2) För dem som passerat tonåren finns gymnasieutbildning inom ramen för kommunal vuxenutbildning. (3) 1971 infördes gymnasieskolformen i Sverige genom att de tidigare skolformerna gymnasium, fackskola och yrkesskola slogs samman då 1970 års läroplan (Lgy 70) trädde i kraft. (4) I vardagligt språkbruk används ordet gymnasium fortfarande om gymnasieskolan. (5) Studentexamen – det vill säga den formella prövningen efter gymnasiestudier för att avgöra höghet till universitetsstudier – avskaffades 1968 i Sverige, men har fortsatt att användas både som begrepp och som firande av avslutade gymnasiestudier. (6) Många gymnasieutbildningar är huvudsakligen förberedande för högre studier, men det finns även de som huvudsakligen är yrkesutbildningar. (7) Slutbetyg från gymnasieskolan krävs för tillträde till högskoleutbildning. (8) Detta kriterium kallas också högskolekompetens. (9) Alla gymnasielinjer ger grundläggande behörighet för högskolestudier. (10) Från 1 juli 1994 reglerades den svenska gymnasieskolan av Lpf 94 (läroplan för de frivilliga skolformerna).

Tabell 6: Undersökningens tre texter.

Den genomgående textkällan är i likhet med många andra implementationer av frågebesvarande system och frågegenerering för svenska och internationellt, den fria encyklopedin *Wikipedia*.

5.2 Undersökningsdeltagarnas karakteristik

Efter utskick med förfrågan om deltagande till studenter genom Göteborgs universitets lärplattform, *GUL*, och genom personliga kanaler slutade deltagarantalet på 19 anonyma personer med fullständigt genomförd undersökning. De ofullständigt genomförda undersökningarna användes inte.[21] Undersökningen gav därmed 171 autentiska frågor från mänskliga läsare, 57 frågor för varje kort textstycke.

Antal fullgjorda deltagarundersökningar:	19
Totalt antal frågeinstanser[22] (med svarsmarkeringar):	171
Antal frågeinstanser (med svarsmarkeringar) per deltagare:	9
Antal frågeinstanser per text per deltagare:	3
Totalt antal frågeinstanser (med svarsmarkeringar) per text:	57

Tabell 7: Data rörande frågor

Det är okänt huruvida undersökningsdeltagarnas ålder eller kön kan ha någon betydelse i detta sammanhang. Om så skulle vara fallet kan förmodligen även andra aspekter påverka.

Undersökningsdeltagarnas könsfördelning:	Kvinnor: 5 (26,3 %) Män: 14 (73,6 %)
Angiven ålder vid tillfället (år):	{19, 20, 22, 23, 23, 25, 32, 33, 35, 35, 36, 36, 36, 40, 41, 42, 45, 47, 67}
Undersökningsdeltagarnas ålder (genomsnitt och median):	$\bar{x}$: 34,6 år $\tilde{x}$: 35 år

Tabell 8: Data rörande studiedeltagarna

[21] Deltagande i undersökningen var troligen relativt krävande. Genom webbenkätsystemet kunde iakttas att många påbörjade undersökningar avbröts.

6 Användarundersökningens resultat

Denna undersöknings empiriska resultatdata har från början ansetts vara värdefulla och intressanta, nästan oavsett vilket utfall de skulle visa på. Resultaten har visat sig kunnat ge stoff åt studiens huvudfråga om de autentiska frågornas likheter eller skillnader gentemot den mängd frågor som skulle kunna produceras rent mekaniskt utifrån de aktuella texterna.

Detta kapitel redovisar undersökningens kvantitativa resultat. Dessa kopplas till i nästa kapitel till undersökningens primära frågeställning med arbetshypotes och till de sekundära frågeställningarna. Detta kapitel pekar också ut några oväntade informationssökande strategier som påträffades det autentiska insamlade materialet.

6.1 Redovisning av undersökningens data: respondenternas svar

Detta första avsnitt innehåller en redovisning av frågorna. De visas först utifrån texternas textmeningar – dvs. vid de segment som besvarar dem, enligt studiedeltagarna. Vidare visas resultatfrågornas fördelning över frågetyp, enligt tidigare beskrivningar.

Huvudfrågan för detta arbete är alltså i vilken mån de autentiska frågorna från användarna ingår bland de som skulle kunna skapas automatiskt med befintlig metod för frågegenerering i sitt nuvarande utförande, eller förbättrad med beskrivna utökningar.

Denna frågeställning har hanterats genom att mängden frågeinstanser (171 stycken) har analyserats. Varje fråga har bedömts, som tillhörande den genererbara kategorin, eller inte.

Det omedelbara utfallet visar att endast 33 av 171 (19,3 %) av frågeinstanserna[23] i denna undersökning är inom direkt räckhåll för den aktuella prototypversionen för svensk text. De frågeinstanser som räknas som inom räckhåll eller nära (direktderivat plus vissa lexikala och syntaktiska) utökningar är markerade med gult nedan. Dessa frågeformuleringar är just nu utom räckhåll, men skulle i vissa fall kunna skapas med en mer utvecklad parsning. För att räknas som inom räckhåll godtogs smärre skriv- och stavningsvariationer. Däremot har bedömningen försökt vara rätt hård i fall som attributfrågor, när det bedöms att det inte riktigt vore möjligt idag.

Analys av attribut och andra satsnivåer (bisatser) skulle fånga in fler frågor. Den viktigaste slutsatsen av resulatet är dock den stora andel frågor som konstruerats på helt annorlunda vis. Se vidare nedan.

[23] Terminologiskt behövs här neologismer för att kunna beskriva och diskutera resultatet. Eftersom den frågemängd som samlas in innehåller samma fråga flera gånger skulle det bli flertydigt att tala om hur många frågor som samlats in. En frågeinstans kommer här att betyda enskild förekomst av fråga. Anledningen till behovet är att i resulterande data förekommer ibland precis samma fråga, ställd av olika undersökningsdeltagare: det innebär alltså flera *frågeinstanser* men en fråga/frågeformulering. Detta är analogt med hur lexikografer m.fl. måste skilja mellan antal types (unika ord) och tokens (ordförekomster) när de behandlar ’antal ord i en text.’ Som exempel fanns i den resulterande mängden av fem frågeinstanser av Hur många invånare har Visby?

Visby	
(1) Visby är en tätort som är centralort i Gotlands kommun och residensstad i Gotlands län.	Vad heter centralorten på Gotland? I vilken kommun ligger Visby? I vilken kommun och vilket län ligger Visby? I vilken kommun är Visby centralort? I vilket län ligger Visby? Vad heter centralorten i Gotlands kommun? Var ligger Visby? Vilken är centralorten i gotlands län?
(2) Visby är även stiftsstad i Visby stift.	**Vilken är stiftstaden i visby stift?**
(3) Visby har 22 593 invånare (2010-12-31).	*Hur många innevånare har Visby?* *Hur många invånare hade Visby i slutet av år 2010?* *Hur många invånare hade visby 2010?* *Hur många invånare har Visby?* *Hur många invånare har Visby?* *Hur många invånare har Visby?* *Hur många invånare har Visby?* *Hur många invånare har Visby?* Hur stort är Visby, hur många invånare har staden? hur många bor i Visby
(4) Den medeltida hansestaden Visby är utan tvekan den bäst bevarade medeltida staden i Skandinavien och är sedan 1995 med på Unescos världsarvslista.	Hur lång är den ringmur som omger den medeltida staden Visby? 1995 hamnade Visby på Unescos världsarvslista, varför? Hur he degrades Visby 1995? När kom Visby med på världsarvslistan? När togs Visby med på Unescos världsarvslista? När togs Visby upp på Unescos världsarvslista **Sedan när är Visby uppsatt på Unescos världsarvslista?** **Vilken är den bäst bevarade medeltida staden i scandinavien?** **4f_Vilken är den bäst bevarade medeltida staden i Skandinavien?**
(5) Bland de mest anmärkningsvärda historiska lämningarna är den 3,4 km långa ringmuren som omger staden och dess gamla kyrkoruiner.	Hur lång är Visby ringmur Hur lång är Visbys medeltida ringmur? Hur lång är Visbys ringmur? Hur lång är den ringmur som omger staden? Nämn en anmärkningsvärd historisk lämning? Vilken av Visbys historiska lämningar är 3,4 km lång? vad är något av det mest anmärkningsvärda med staden? Är Visbys ringmur längre än en mil?
(6) Visby är ett populärt resmål under sommaren och tar emot tusentals turister varje år.	Vilken årstid lockar Visby flest turister? hur många Turister utsätts ön för på sommaren?
(7) Visby är säte för Högskolan på Gotland.	Finns det ett universitet på Gotland? Finns det ett universitet i Visby I vilken stad ligger Högskolan på Gotland Vad heter universitet I Visby?
(8) Visby kallas 'rosornas och ruinernas stad'.	Vad brukar Visby kallas? **Vad kallas Visby?** **Vad kallas Visby?** Visby kallas även?
(9) De äldsta fynden på platsen för dagens Visby är vad som tolkats som 'strandbodar', med C14-metoden daterade till 700-900-talen e Kr.	Hur gamla är de älsta fynden i Visby Hur daterades standbodar? Hur gamla är de äldsta arkeologiska fynden i Visby? Hur gamla är strandbodarna? Med vilken metod har man kunnat datera åldern på Visbys strandbodar? Till vilka århundraden dateras de strandbodar som hittats på platsen för dagens Visby? Ungefär vilket århundrade har man daterat de äldsta fynden i Visby till? **Vad är de äldsta fynden i Visby?** **Vilka är de älsta historiska lämningarna kring Visby?** **Vilket är de äldsta fynden i Visby?**

(10) Längs vikingatidens gotländska kust fanns en rad hamnar med gravfält och viss bebyggelse.	Vad är det mest anmärkningsvärda med vikingatidens gotländska hamnar?
Julio Baghy	
(1) Julio Baghy (ungerska Baghy Gyula), född 13 januari 1891 i Szeged, död 18 mars 1967, var en ungersk författare och skådespelare.	När är Julio Baghy född) Från vilken land kom Julio Baghy? Från vilket land kom Julio Baghy? Från vilket land kom Julio Baghy? Hur gammal blev Julio Baghy? Hur gammal blev den ungerske författaren Julio Baghy? När dog Julio Baghy? När föddes Julio Baghy? När föddes Julio Baghy? När föddes Julio Baghy? När föddes Julio Baghy? När föddes han och hur gammal blev han? **Vem var Julio Baghy?** **Vem var Julio Baghy?** **Vem var Julio Baghy?** Vilket land kom Julio Baghy från? hur kallas han på ungerska? när föddes Julio Baghy 1f Vad jobbade Julio Baghy med? 1f Vilket var Baghys hemland?
(2) Han lärde sig esperanto 1911 och började arbeta för esperantorörelsen under en sexårig krigsfångenskap i Sibirien.	2, 10 Vilken rörelse blev Julio sedermera en av de viktigaste ledarna för? I hur många år levde Julio Baghy i krigsfångenskap? 2-10 Var han med i kriget? 2. Men även 2-10: Vad hade han för yrke? Hur många år spenderade Julio Baghy i rysk krigsfångenskap? **När lärde Julio Baghy esperanto?** **När lärde sig Julio Baghy esperanto?** Vilket år lärde sig Julio Baghy esperanto? **när lärde han aig Esperanto?** 2f Med vilket språk förknippas Julio Baghy? 2f Under vilka förhållanden arbetade Baghy för esperantorörelsen?
(3) Han var en framstående esperantoaktivist och lärare, och under en tid vice ordförande för Esperantoakademin.	Vad är Julio Baghy känd för? Vilket språk blev Julio Baghy en aktivist för? Vilket yrke hade han? vad jobbade han med 3f Vilket språk är han en känd beivrare för? 3f Vad är Julio Baghy mest känd för?
(4) Baghys far var skådespelare och hans mor teatersufflör.	Vad arbetade Julio Baghys föräldrar med? Vad arbetade hans föräldrar som? Vilket yrke hade Baghys mor? Vilket yrke hade Julio Baghys far? Vilket yrke have Julio Baghys fader?
(5) Efter skolgången blev även han skådespelare och regissör vid olika teatrar.	Med vad arbetade Baghy efter skolan? 5f Hur inledde Julio Baghy sitt yrkesliv efter skoltiden?
(6) Kriget avbröt hans karriär och han tvingades tillbringa sex år utanför sitt hemland i rysk krigsfångenskap.	Hur länge satt Julio Baghy i rysk fångenskap? Hur länge satt han i rysk fångenskap? **Vad avbröt hans karriär?** 6f, Var satt han under fånge under sex år?
(7) Redan under hans ungdom publicerades många av hans dikter och noveller i ungerska tidskrifter.	

(8) 1911 lärde han känna esperanto, och dess inre idé (interna ideo) lockade honom omedelbart.	**När lärde Julio Baghy känna Esperanto?** **När lärde han känna esperanto?** Vad hette esperantorörelsen vilket han ledde I ungerska?
(9) Hans omfattande engagemang för esperanto började redan i fånglägret i Sibirien, där han ledde flera kurser för människor från olika länder.	Var satt han i fängelse? Var började Baghy hålla kurser i Esperanto? 9f10 Vad hade Julio Baghy för roll i esperantorörelsen?
(10) Sedan han återvänt till Ungern efter kriget blev han en av de viktigaste ledarna för esperantorörelsen: han höll många kurser på olika nivåer, ledde Esperanto-Rondo Amika, arrangerade litterära kvällar och så vidare.	Nämn en sak som Julio Baghy gjorde när han var en av de viktigaste ledarna för esperantorörelsen. Vad gjorde Baghy efter kriget? Vad ägnade Julio Baghy sig åt när han återvände till Ungern?
Gymnasieskola	
(1) En gymnasieskola är en avgiftsfri och frivillig sekundärutbildning i Sverige för ungdomar som har gått ut grundskolan.	Ska man betala för att gå på gymnasiet i Sverige? Hur gammal är man när man börjar gymnasiet? Hur mycket betalar en genomsnittslig gymansiestudent i kursavgifter per läsår i Sverige 2013? Måste alla svenskar ha gymnasieutbildning? Måste man gå gymnasiet? Vad kostar det att gå i gymnasiet i Sverige? Vad kostar det att läsa på gymnasiet i Sverige? Vad kostar det att studera på gymnasienivå i Sverige? Vad kostar gymnasieskolan? Vad kostar svensk gymnasieskola? **Vad är en gymnasieskola?** **Vad är en gymnasieskola?** **Vad är en gymnasieskola?** *Är gymnasiet frivilligt och hur mycket kostar det? (täcks till hälften)*
(2) För dem som passerat tonåren finns gymnasieutbildning inom ramen för kommunal vuxenutbildning.	Hur kan vuxna komplettera gymnasieutbildning? Kan vuxna studera på gymnasiet? Vad finns det för alternativ för de som inte genomgått gymnasiet? Vem håller i gymnasieutbildning för dem som passerat tonåren?
(3) 1971 infördes gymnasieskolformen i Sverige genom att de tidigare skolformerna gymnasium, fackskola och yrkesskola slogs samman då 1970 års läroplan (Lgy 70) trädde i kraft.	**När infördes gymnasieskolformen i Sverige?** **När infördes gymnasieskolformen i Sverige?** **När infördes gymnasium i Sverige?** När slogs de tidigare skolformerna gymnasium, fackskola och yrkesskola samman? Vad innebar 1970 års läroplansändring? Vilka skolformer slogs samman till gymnasium? Vilka tre tidigare skolformer slogs samman 1970 för att skapa den moderna gymnasieskolformen 1971? Vilket år infördes gymnasieskolformen i Sverige genom att de tidigare skolformerna gymnasium, fackskola och yrkesskola slogs samman? **när infördes gymnasieskolformen**
(4) I vardagligt språkbruk används ordet gymnasium fortfarande om gymnasieskolan.	
(5) Studentexamen – det vill säga den formella prövningen efter gymnasiestudier för att avgöra höghet till universitetsstudier – avskaffades 1968 i Sverige, men har fortsatt att användas både som begrepp och som firande av avslutade gymnasiestudier.	Vad kallas den examnen elever får, när man har klarat av gymnasiestudierna? Hur går en studentexamen numera till i Sverige? När avskaffades den formella prövningen (examinationen) efter genomgångna gymnasiestudier? **När avskaffades studentexamen i Sverige?** **När avskaffades Studentexamen?** Vad innebar studentexamen fram till år 1968 då dess ursprungliga innehåll avskaffades? Vad var det man avskaffade 1968 och vad innebar det?

(6) Många gymnasieutbildningar är huvudsakligen förberedande för högre studier, men det finns även de som huvudsakligen är yrkesutbildningar.	Är alla gymnasieutbildningar högskoleförberedande? Är gymnasium högskoleförberedande? 6f Vad finns det för skäl att skaffa sig en gymnasieutbildning?
(7) Slutbetyg från gymnasieskolan krävs för tillträde till högskoleutbildning.	**Krävs det slutbetyg från gymnasieskolan för tillträde till högskolestudier?** Vad innebär högskolekompetens? **Vad krävs för att få studera vid en högskola?** **Vad krävs för att idag få tillträde till högskoleutbildning** **Vad krävs för tillträde till högskolan?** **Vad krävs för tillträde till högskoleutbildning?** Vad är högskolekompetens? vad krävs för att bli behörig till gymnasiet? 7f Vad betyder begreppet högskolekompetens? 7f8f9f Vad är högskolekompetens? 7f_Vad omfattar det kriterie som kallas 'högskolekompetens' från gymnasiet?
(8) Detta kriterium kallas också högskolekompetens.	Vad är högskolekompetens?
(9) Alla gymnasielinjer ger grundläggande behörighet för högskolestudier.	**Ger alla gymnasieutbildningar högskolekompetens?** Vilka av de gymnasieutbildningar som finns tillgängliga idag ger grundläggande behörighet för högskolestudier? Vilka gymnasielinjer ger behörighet till högskola? Vilka gymnasielinjer ger grundläggande högskolebehörighet? Vilka linjer ger grundläggande kompentens för högskolestudier vad krävs för behörighet för högskolestudier?
(10) Från 1 juli 1994 reglerades den svenska gymnasieskolan av Lpf 94 (läroplan för de frivilliga skolformerna).	När trädde Lpf 94 i kraft? Vad reglerar den svenska gymnasieskolan?

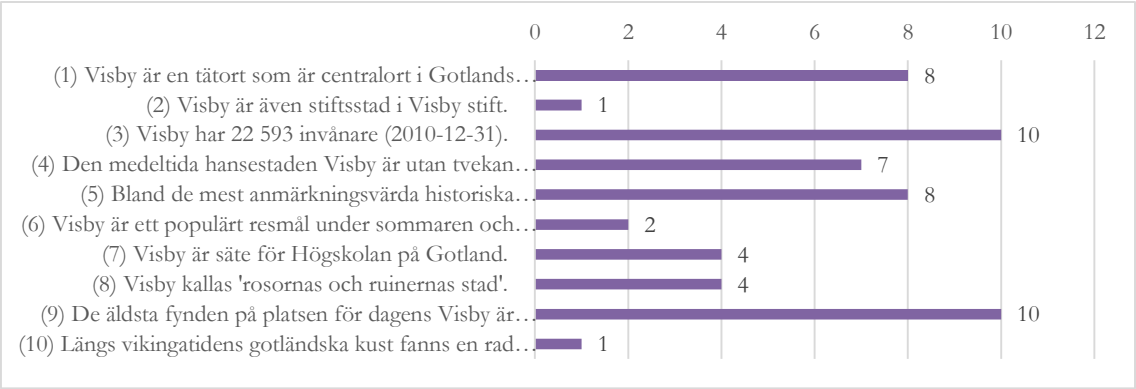
Tabell 9: Texterna med angivna tillhörande frågor.

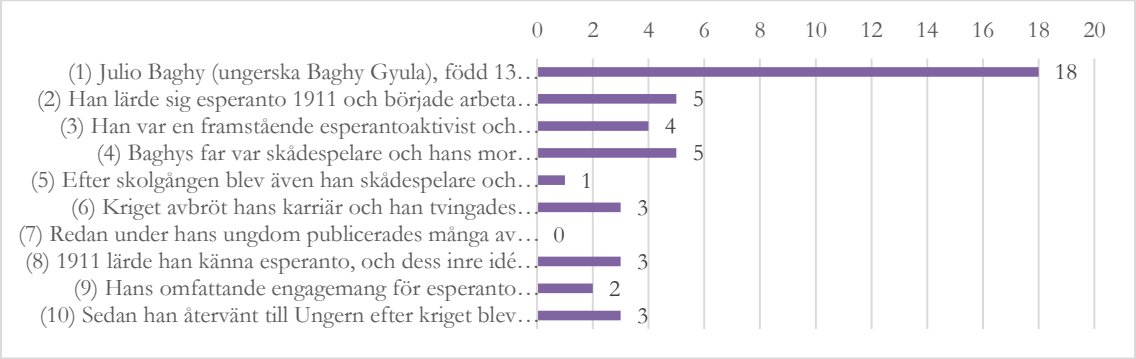
Som nämndes finns en del frågeinstanser något bortom det som kan genereras för närvarande. En sådan bedömning är svår att göra exakt men ytterligare uppemot tio frågor (ca 5,8 %) bedöms kunna täckas in med en något förbättrad frågegenerering. Några attributfrågor, som de kursiverade till Visby-textmening 3, ligger t.ex. inom någorlunda räckhåll.

6.2 Undersökningens sekundära frågeställningar: vilka textsegment utgör svar?

Med vissa undantag, som frågan om invånarantalet i Visby, är frågefördelningen relativt spridd över texternas ingående textmeningar.

Nedanstående markering av svarstextmening (eller första svarsmening i förekommande fall) visar att *inledande textmeningar* ofta är omfrågade (8, 18 resp. 14 frågor). Om fördelningen av frågeinstanser över textmeningar vore helt jämn skulle varje textmening omfrågas 5,7 gånger.





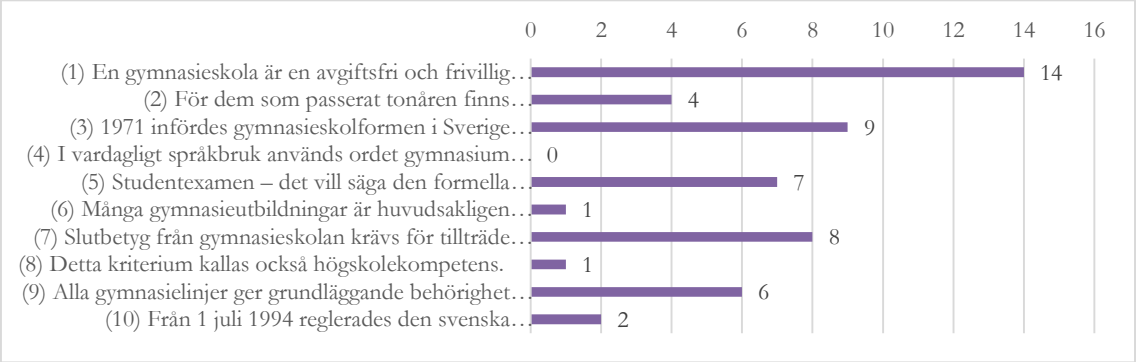
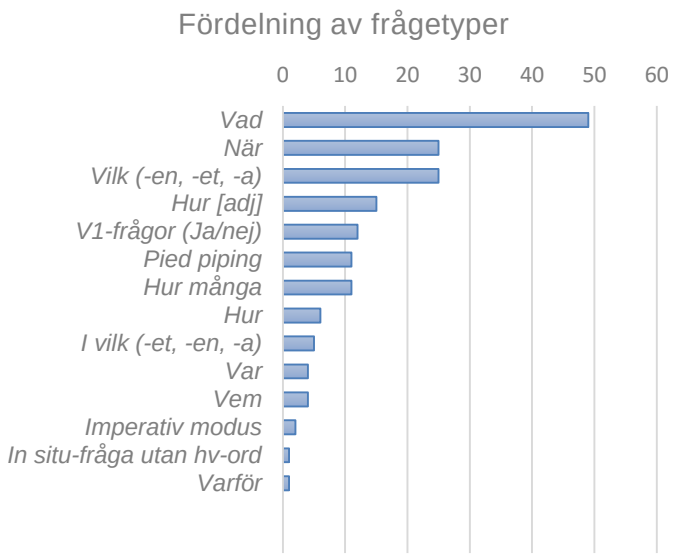


Diagram1, 2, 3: Antal frågor per textmening

Inom teori om språklig *informationsstruktur* studeras hur syntax på olika sätt kan påverka informationsprominens (i nordiska språk, bl.a. genom spetsställning, se kapitel 2). Medan prosodin markerar vad som är betonat i talspråk – finns liknande markörer i syntaxen som kan lyfta fram vad skribenten ser som viktigt i en skriven text. Det ska sägas att forskningsfältet är relativt livfullt, men undersökningar där testpersoner ombeds markera informationsmässigt framträdande dela visar låg *co-annotator agreement*, dvs. människor är olyckligtvis inte speciellt eniga om vad dessa begrepp betyder, alternativt vad som är informationsprominent i en text (Ritz, Dipper, & Götze, 2008).

6.3 Undersökningens sekundära frågeställningar: frågornas kvantitativa fördelning



Vad	49
När	25
Vilk (-en, -et, -a)	25
Hur [adj]	15
V1-frågor (Ja/nej)	12
Pied piping	11
Hur många	11
Hur	6
I vilk (-et, -en, -a)	5
Var	4
Vem	4
Imperativ modus	2
In situ-fråga) utan hv-ord)	1
Varför	1

Diagram 4: Fördelning av frågetyper, innehållande några oväntade konstruktioner.[24] Attributfrågetypen *Hur många* tillhör de som också skulle kunna täckas med relativt enkla medel.

Textinnehållet kan givetvis styra de frågor som ställs varför frågetypsfördelningen inte får övertolkas.

När undersökningen konstruerades och uppgiften formulerades för undersökningsdeltagarna var tanken att instruktionerna skulle vara enkla och korta men samtidigt entydiga. Det förutsattes att den indata som skulle ges: markering av en textmening där svaret på fråga fanns – eller när svaret fanns på flera håll – numret för den första svarsmeningen följt av ett f (t ex 5f) skulle fungera som heltäckande alternativ. Bland respondenternas svar finns dock intressanta avvikelser när det gäller att markera svar som hämtas in från flera textmeningar. Flera respondenter har markerat på ett annorlunda sätt: t ex ”3f, 9”. Någon har markerat ”2, 10”.

[24] Pied piping innebär i grammatiska termer att en prepositionsfras omfrågas med komplementsdelen (rektionen) omgjord till *hv*-led. Exempel: *Med vad arbetade Baghy efter skolan?*Imperativ: *Nämn en anmärkningsvärd historisk lämning? In situ*-fråga utan *hv*-ord (elliptisk konstruktion): *Visby kallas även?*

7 Diskussion och slutsats

Detta kapitel gör en analys av användarundersökningens resultat. Något om de naturliga frågornas egenskaper jämfört med de genererade nämns. I detta avslutande kapitel sammanfattas också det som uppfattas som studiens huvudslutsatser.

I den vetenskapliga teoribildningen innebär hypotesframställning inför en empirisk undersökning, tillsammans med en vederlagd arbetshypotes en grogrund för något nytt. *Hypotesen*, den föreslaget allmänna lagen (Patel och Davidson, 1994) revideras med tanke på de mätvärden som erhållits. I det strängt empiriskt betonade tillvägagångssättet framtas en ny, reviderad hypotes och därefter vidtar nya empiriska försök med denna föreslagna lags förutsägelser mot ny data.

Frågeställningen här har varit huruvida det i kvantitativt perspektiv finns en tydlig relation mellan autentiska frågor och de som kan åstadkommas på konstgjord väg. De mätvärden som har inhämtats genom undersökningen har givit ett användbart resultat och pekar på ett numerärt förhållande av ett visst slag, vilket beskrivits i föregående kapitel. Poängen med den aktuella undersökningen har varit att se i vilken storleksordning som sambandet råder. Resultaten är oavsett vidare resonemang ett användbart motbevis till idén om genomgående mekanisk genererbarhet med aktuell ansats för svenska. Denna s.k. arbetshypotes vederläggs i denna undersökning genom en samling motexempel.

Åter till inledningen av denna uppsats:

> Visby har 22 593 invånare. Bland de mest anmärkningsvärda historiska lämningarna är den 3,4 km långa ringmuren som omger staden och dess gamla kyrkoruiner. Visby är ett populärt resmål under sommaren och tar emot tusentals turister varje år. Visby är säte för Högskolan på Gotland. Visby kallas 'rosornas och ruinernas stad'.
>
> Från *Visby* (Svenska *Wikipedia*), modifierad.

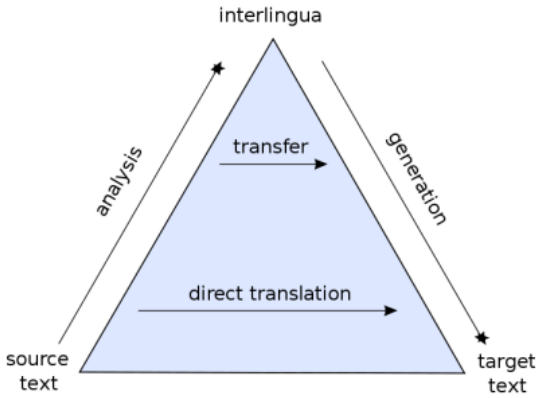
a) Vad kallas Visby?
b) Hur många bor i Visby?
c) Är Visbys ringmur längre än en mil?

Fråga a) är vad som här kallats *direktderivat* och alltså tillhör det genererbara, b) innebär en *omformning av både syntax och lexikon (verbkonstruktion)*, se vidare nedan. Den hamnar utanför täckningen, liksom c) som kräver en ytterligare sorts semantisk analys. I undersökningen har alltså endast omkring 20 % av frågeinstanserna räknats till den första gruppen.

7.1 Frågor ur läsarens minne

En definitiv slutsats att dra från undersökningen är att mänskliga läsare av en text inte utgår från den skrivna syntaxen och skapar direktderivat på ett ”mekaniskt” sätt, när de ombeds skriva relevanta frågor. När relevanta frågor ställs om en text utgår läsaren som nu är frågeställare från sin förståelse av texten. När läsaren utgår från sin förståelse av texten (och t.ex. ombes att återberätta innehållet, eller, som här, att ställa frågor om det) formulerar läsaren innehållet självständigt och med egna ord. Detta sker inte genom en trogenhet mot den faktiska textformuleringen (som de flesta läsare troligen snabbt glömmer) utan utgår från en förståelse, en mental representation av innehållet som läsaren klär med sina egna ord. (Även om källtexten faktiskt visas i klartext för läsaren hela tiden här.)

Här finns en analogi med hur automatisk översättning (maskinöversättning) fungerar i jämförelse med äkta, mänsklig, översättning, vilket innebär att översättaren först läser och förstår texten. Innehållet kläs i ord på målspråket. Hur innehållet/betydelsen i mellanläget representeras (eventuellt ”språklöst”) hos läsaren är naturligtvis svårt att veta, men att denna mellanform existerar gör det annorlunda jämfört med fungerande maskinöversättningssystem som oftast inte har någon sådan språkoberoende tanke- eller betydelsenivå. Detta tankesätt brukar illustreras mycket väl med hjälp av Vauquois triangel. Ju högre nivå på triangeln som involveras, desto djupare är analysen och ju bättre är möjligen översättningen.



Figur 14: *Vauquois triangel*[25] används för att illustrera hur *översättning* mellan två språk kan ske genom olika nivåer. En mänsklig översättare analyserar i läsningen källspråkets uttryck och kommer fram till den ”högsta” analysnivån. På denna nivå där innehållet eventuellt representeras ”språklöst” eller ”språkoberoende” kläs meddelandet i både ny syntaktisk form och nya ord på målspråket. En maskinell översättning går så att säga inte hela vägen till denna språkoberoende nivå – det som i så fall skulle kallas en *interlingua*-representation.[26] Direktöversättning brukar visas som exempel på den absolut sämsta typen av översättning. Den sker ord för ord och utan hänsyn till ordföljdsskillnader i språkpar.

På samma sätt som en översättare i något läge kan lagra betydelsen från källspråket vid en översättning innebär så har den mänskliga frågeformuleraren utgått från minnet eller förståelsen på ett sätt som skiljer sig från de mekaniskt producerade direktfrågorna.

Att de autentiska frågorna inte sammanfaller med direktfrågorna är ett förhållande som liknar det faktum att maskinöversättning bara till liten del sammanfaller med en mänsklig översättning. I fallet översättning är detta dock inte så nedslående eftersom det antas finnas många korrekta översättningar.

Om den omfrågade informationen inte är den som omfrågas av direktfrågorna, vad är den då? Undersökningen har tydligt visat att de frågor eller frågeformuleringar som ställs naturligt inte tillhör gruppen direktderivat. Men finns inte svaren på frågorna ändå bland de funktionella led som går att identifiera automatiskt. De led som utgör svar på läsarnas egna frågor finns i själva verket visst ofta bland de satsled som omfrågas. Själva frågorna är dock inte direktderivat eller sådana som är uppbyggda av de satskonstruktioner (ord och syntax) där svarsledet ingår.

I undersökningen framträder, i linje med det ovanstående resonemanget, en tydlig tendens rörande den uttryckliga informationen i text och de frågor som blir resultatet.

[25] Bernard Vauquois var på 1970-talet en inflytelserik forskare inom maskinöversättning vid universitet i Grenoble.
[26] Interlingua ska inte förväxlas med det konstgjorda språket (av typ esperanto) som också bär detta namn.

(4) Baghys far var skådespelare och hans mor teatersufflör.

↑

4_Vilket yrke hade Julio Baghys far?

Gång efter gång syns hur den fråga som ställs och som besvaras av textmeningar – som den ovanstående – inte är ett frågeuttryck som automatisk frågegenerering skulle ge. Från textmeningen *Baghys far var skådespelare och hans mor teatersufflör'* väljer användaren inte att ställa vad som skulle vara den automatgenererade motsvarigheten (*Vad var Julio Baghys far?)* Istället återfinns frågan *Vilket yrke hade Julio Baghys far?* Denna 'parallellfråga' som besvaras av samma information (skådespelare) ser istället ut att ha sitt upphov i en annan sats *(Julio Baghys far hade yrket skådespelare).* Denna andra sats som så att säga har använts av frågeställaren innehåller en innehållsmässig transformation; Att vara något X (där X är ett yrke) till att ha yrket X. (Frågan gäller från början predikatet i satsen till att ha transformerats till en fråga där ett attribut omfrågas.)

I resultatet från undersökningen blir det tydligt hur stor andel av de ställda frågorna som inte är av den explicita typ som frågegenereringen åstadkommer, utan som istället innehåller denna avancerade semantiska transformation. Detta är en av de mest tydligaste slutsatserna i detta arbete. Typen av frågegenerering som krävs innehåller den mellanform av både lexikal och syntaktisk omvandling som tillskrevs Lin och Pantel (2001) ovan. Dessa författare konstruerar både egna och använder en algoritm för att ta fram sådana likvärdiga uttryck i engelska.

X is author of Y	***X manufactures Y***
Y is the work of X; *X is the writer of Y;* *X penned Y;* *X produced Y;* *X authored Y;* *X chronicled Y;* *X wrote Y*	*X makes Y;* *X produce Y;* *X is in Y business;* *Y is manufactured by X;* *Y is provided by X;* *Y is X's product;* *Y is product from X;*

Tabell 10: Exempel på de likhetsförslag som framträtt i Lin och Pantels arbete (2001).

Ovanstående tabell visar ganska många ganska enkla ordutbyten. Det som framför allt avses är fall där verbkonstruktion helt ändras tillsammans med leden, som *X manufactures Y – Y is product from X.*

7.2 *Inbjuder undersökningens upplägg undersökningsdeltagarna att ställa andra frågor än direktderivat?*

I den aktuella undersökningen bads bara användare att ställa relevanta frågor som besvaras av en text. Det finns inga instruktioner som säger att användaren måste ställa en mer raffinerat konstruerad fråga än den typ som här benämns direktderivat. Resultatet visar emellertid att de naturliga frågor som ställs av användare till övervägande del varken är direktderivat eller lite mer indirekt deriverade frågeformuleringar. Kan det vara så att undersökningens upplägg, trots att det inte förekommer i instruktionerna, ibland har föranlett deltagarna att konstruera frågor som så att säga kräver en djupare analys, för att detta skulle vara någon sorts norm för frågeställande?

I somliga sammanhang som högskoleprovets LÄS-uppgift nedan,[27] verkar det nog orimligt att uppgiften skulle besvaras så enkelt som att frågan skulle vara ett direktderivat; men det är förstås en speciell kontext. Provkonstruktören vill vinnlägga sig om läsarens förståelse av texten.

[27] Övningsprov, LÄS, läsförståelse, Högskoleprovet 2010

När det gäller att skriva lite längre texter gör de flesta skribenter först någon slags förplanering, antingen i sitt huvud, på papper eller på skärmen. De gör också en slutgenomläsning och kollar att allt hänger ihop.

Men kanske mer intressant med den skickliga skribenten är att hon kan bolla med alla de här delarna under hela skrivprocessen. Hon inte bara skriver ned de idéer hon hade innan texten börjar, utan genererar nya idéer utifrån sitt eget skrivande, anpassar texten och planerna till dem och utvärderar och ändrar hela tiden medan hon skriver för att verkligen anpassa till läsaren och målet med skrivandet. Man kan säga att hon omvandlar både sina idéer och sin text allteftersom hon skriver den.

Nybörjarskribenten däremot har fullt upp med huvudsakligen två saker. Det kanske absolut svåraste är att kommunicera med en icke närvarande mottagare. Innan man lär sig skriva innebär kommunikation med andra människor oftast ansikte-mot-ansikte-kommunikation och ibland kanske telefonsamtal. I båda de här situationerna är båda konversationsparterna närvarande vid samma tidpunkt och det gör att den som talar omedelbart får återkoppling på om den som lyssnar har hört, förstått och reagerat på vad man just har sagt. Man använder också ett språk som är anpassat till att man mer eller mindre befinner sig i samma kontext och troligen känner varandra. Man kan peka och referera till saker i omgivningen utan att behöva förklara hur de ser ut och var de står osv. När man ska lära sig skriva måste man lära sig att kommunicera med någon som inte är närvarande och som man kanske inte ens känner. Det här kan ofta påverka ens ordval och val av grammatiska strukturer. Den andra biten som nybörjarskribenter ofta har svårt med är inte särskilt förvånande, nämligen de ibland kallade ”mer mekaniska” delarna av skrivande, alltså att forma bokstäver, hitta bokstäver på tangentbordet och sist men inte minst att stava. Att detta är svårt är de flesta nybörjarskribenter väl medvetna om och om man frågar en nybörjarskribent vad skrivande är säger de flesta något i stil med att stava eller hitta bokstäverna. Så länge man inte automatiserat de processerna har de en tendens att ta upp större delen av skribentens kognitiva resurser och det finns små möjligheter att lägga kraft på något annat. Det här beror enligt hypotesen om kognitiv kapacitet på att vi inte har obegränsade kognitiva resurser att ta till, utan för varje uppgift vi ska genomföra (t ex spela ett datorspel eller skriva en text) finns det bara en viss begränsad mängd kognitiv kapacitet att ta till och om en delprocess tar upp väldigt mycket kraft blir det helt enkelt inte så mycket över till den andra.

Åsa Wengelin

Uppgifter

15. Vad utmärker enligt texten skrivprocessen hos en skicklig skribent?

A Grundligt förarbete och noggrann planering.
B Fortlöpande bearbetning och korrigering.
C Förmåga att anpassa språket till ämnet.
D Att känna vad läsaren föredrar.

16. Vad är huvudpoängen i textförfattarens avslutande resonemang om kognitiva resurser?

A Att den kognitiva kapaciteten varierar mellan individer.
B Att skrivprocessen fordrar omfattande kognitiv kapacitet.
C Att den kognitiva kapaciteten efterhand automatiseras allt mer.
D Att automatiserade processer frigör kognitiv kapacitet.

7.2 Kontextlösa frågor jämfört med autentiska frågor med kontext

Finns det andra skillnader mellan frågor som skapas automatiskt och de som verkliga användare ställer? Det är möjligt att i efterhand spåra en sannolik skillnad som hänger samman med kontext. De automatiskt genererade frågorna som utgår från varje textmening där den står. Sammanhanget för ursprungsmeningen återskapas inte i automatgenererade frågor. Det betyder t.ex. att pronomen osv. finns med outredda. Ett tänkbart direktderivat som ”Varåt gick han då?” är en så kontextlös att den kanske aldrig skulle ställas till ett system med stor databas. Den autentiska frågan som ställs av läsare däremot – som inte direkt beror på ordalydelsen, utan betydelsen – blir generellt förbättrad och förtydligad av den frågeställaren så att all nödvändig information finns med. Användaren som ställer en fråga utan att den språkliga kontexten finns där, måste i en normal situation bygga in en temporär kontext i frågan för att göra den entydig och begriplig. Detta exemplifieras i många av de insamlade frågorna.

7.3 Att fråga eller att icke fråga

Frågebesvarande system kan placeras in i en djupare kontext för att klargöra vilken roll de kan uppbära för en informationssökande individ (t.ex. inom en organisation). Om ett frågebesvarande system verkligen ska kunna ses som *tillförlitligt* (vilket den närliggande termen expertsystem t.ex. antyder), hur ser möjligheterna ut? Den bakomliggande frågan är: Går det att *lita* på en godtycklig text, och i förlängningen, på systemtypen?

I ett vetenskapsteoretiskt perspektiv förekom under 1900-talet en kritik rörande somliga discipliners förmenta objektivitet. I den positivistiska andan där forskningsämnen som matematik och fysik framhölls som längst komna i korrekt framställd vetenskap (exempel från Beard, 1935,[28] hämtat från Kjørup, 1996) kunde djup skepsis råda rörande värdet av en historikers gärning. Historiker studerade nästan enbart det förgångna, genom spår och inte egna iakttagelser. Forskningen inom historia bedrevs nästan uteslutande genom andras källor – texter – och ofta andrahandskällor. Kunde en sådan disciplins ”kunskap” över huvud taget ha rätten att kallas objektiv, på samma sätt som en kemists nyss personligen utförda experiment?

Frågebesvarande system och söksystem för informationsåtkomst över huvud taget har förutsättningar som liknar historikerns i exemplet – texterna kan innehålla alla tänkbara faktafel (även om de skulle vara nyskrivna). Om texten *ljuger* finns inget att göra. Men här kan även pekas på en positiv aspekt som implementationer av denna form av informationssökning faktiskt har. *PowerSet* som nämndes i inledningen och systemet för svenska Wilhelmsson (2011) har båda egenskapen ’att inte ljuga själv’ inbyggd.’ Naturligtvis kan systemet inte avgöra om texten innehåller lögner: textmängdens försanthållanden är det enda som existerar som kunskapsbas. – Däremot finns en försiktighetsaspekt hos flera implementationer. I den svenska QG-implementationen är tanken att användaren efter ställd fråga (närmare bestämt val av en automatiskt generad fråga) styrs till det segment i texten som givit upphov till frågan (och som förhoppningsvis därmed är ett korrekt svar). I samma anda levererade *PowerSet* som svar en rankad lista med *sannolikt* besvarande segment och lämnade över själva uttydandet till användaren (frågeställaren).

Dessa användargränssnitt riskerar i dessa utföranden inte att anklagas för felaktig information/svar eftersom de i det ena fallet bara ”svarar” när användaren valt en genererad fråga och i det andra fallet enbart levererar en samling statistiskt/heuristiskt rankade segment, vilka bara *eventuellt* besvarar frågan. De är på ett sätt, formen till trots, inte olika vanliga sökströngbaserade

[28] Ur *A noble dream*, 1935

informationssökningssystem, som ju också lämnar t.ex. en lista till webbsidor eller andra dokument som svar, istället för att behöva välja ett enda.

'Modigare' gränssnittsansatser ska alltid leverera precis ett svar och kan exemplifieras av *Watson*. Till denna kategori hör troligen delvis också röststyrda dialogsystem som inte kan ge visuell respons på skärm. Röststyrda telefonsystem – i svenska förhållanden pionjärsystemet i SJ:s resebeställningstjänst 0771-757575 – och numera i mängder av automatiska kundserviceinrättningar visar att en tillräckligt insnävad domän kan ge god funktion, trots de stora svårigheterna inom automatisk röstanalys. Kan dialogsystemet dessutom lära sig det aktuella ämnet under samtalets gång och insnäva tolkningarna ytterligare är det fördelaktigt (Jonson, 2010).

7.4 Ett avslutat kapitel

I en enkel mening, givet formulerad arbetshypotes (se kapitel 3), är utfallet för denna studies undersökningsfråga negativt. Den implementation av frågegenerering för informationsåtkomst i svensk text som framförallt har funnits i åtanke levererar i nuvarande utförande endast en bråkdel av användares verkliga frågeformuleringar. Även med tänkta utökningar faller många av de faktiskt relevanta frågorna/frågeformuleringarna, enligt de mänskliga läsarna, utanför ramarna. Är då komponenten frågegenerering i denna sorts informationssystem ett avslutat kapitel?

Det bör konstateras att denna programkomponent har funnits till för andra ändamål än just informationssystem mot fri text. Lefevre et al. (2009) har exempelvis undersökt automatisk frågegenerering för att producera studentfrågor för läsförståelsetester.

När det gäller studiet av frågor och svarsformer torde det ha en allmän giltighet för informationssystem med naturligt språk även om de inte involverar frågegenerering.

7.5 Slutsatser

- Huruvida autentiska frågor idag skulle kunna skapas av det svenska QG-systemet, undersöktes kvantitativt. Det visade sig att bara omkring 20-25 % av frågeinstanserna antas vara inom nära räckhåll. Kanske kan ytterligare 10 % nås med viss förbättringar för t.ex. attributfrågor.

- En majoritet av de frågeinstanser som samlats in visar prov på en omformning utifrån innehållet inte bara i ord eller grundläggande syntaxtransformationer, utan i konstruktionstyp, inklusive satsverb, roller osv. Detta är en tydlig signal om hur ett steg av mänsklig förståelse gärna helt raderar ursprungsuttrycket och stöper om frågans beståndsdelar från grunden.

- Denna omformulering av satser och/eller frågor på konstruktionsmässiga grunder som beskrivs som nödvändig för att nå fram till de autentiska frågorna exemplifieras tydligt i relaterad forskning för engelska av Lin och Pantel (2001).

8 Referenser

A Dictionary of Philosophical Terms and Names. (n.d.). Retrieved 12 01, 2009, from http://www.philosophypages.com

Ask Jeeves. (n.d.). Retrieved from Ask Jeeves: www.ask.com

Beynon-Davies, P. (2013). *Business Information Systems.* Palgrave Macmillan.

Breitholtz, E. (2010). Enthymematic Inference in Dialogue. *Inference in Dialogue Workshop.*

Carlberger, J., & Kann, V. (1999). Implementing an Efficient Part-Of-Speech Tagger. *Software—Practice & Experience*, 815 - 832.

Converse, T., Kaplan, R. M., Pell, B., Prevost, S., Thione, L., & Walters, C. (2008). Powerset's Natural Language Wikipedia Search Engine. *Wikipedia and Artificial Intelligence: An Evolving Synergy. Papers from the 2008 AAAI Workshop* (p. 67). Chicago, USA: AAAI Press.

Davenport, T. H. (1997). *Information Ecology.* New York: Oxford University Press.

Diderichsen, P. (1946). *Elementær Dansk Grammatik.* Köpenhamn: Gyldendahl.

Ejerhed, E., Källgren, G., & Brodda, B. (2006). *Stockholm-Umeå corpus version 2.0.* Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet.

Ericsson, S. (2006). *Information Enriched Constituents in Dialogue (Doktorsavhandling).* Göteborg: Göteborgs universitet.

Fraurud, K. (1988). Pronoun Resolution in Unrestricted Text. *Nordic Journal of Linguistics 11*.

Grice, P. (1975). Logic and conversation. *Syntax and semantics*.

Harris, Z. (1954). Distributional Structure. *Word*.

Heilman, M., & Smith, N. A. (2009). Ranking Automatically Generated Questions as a Shared Task. *Proceedings of the AIED Workshop on Question Generation.* Brighton.

Holm, L., & Larsson, K. (1980). *Svenska meningar: Elementär språklära.* Lund: Studentlitteratur.

Hultman, T. G. (2003). *Svenska akademiens språklära.* Stockholm: Svenska akademien. Norstedts ordbok distributör.

Jonson, R. (2010). *Information state based speech recognition (doktorsavhandling).* Göteborg: Göteborgs universitet.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing.* Prentice-Hall.

Jörgensen, N., & Svensson, J. (1986). *Nusvensk grammatik.* Malmö: Gleerups.

Kann, V., & Rosell, M. (2005). Free Construction of a Free Swedish Dictionary of Synonyms. *Proceedings of 15th Nordic Conference on Computational Linguistics – (NODALIDA 05).* Joensuu.

Kann, V., & Rosell, M. (2005). Free Construction of a Free Swedish Dictionary of Synonyms. *Proceedings of 15th Nordic Conference on Computational Linguistics (NoDaLiDa 05).* Joensuu.

Kjørup, S. (1996). *Människovetenskaperna.* Studentlitteratur.

Lefevre, M., Jean-Daubias, S., & Guin, N. (2009). Generation of Exercises within the. *Proceedings of the AIED Workshop on Question Generation.* Brighton.

Lin, D., & Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*.

Lyngfelt, B., & Forsberg, M. (2012). *Ett svenskt konstruktikon. Utgångspunkter och preliminära ramar.* Göteborgs universitet, GU-ISS.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. *Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference.* Montreal.

Montague, R. (1970). English as a Formal Language. In B. Visentini, *Linguaggi nella società e nella tecnica.*

Nilsson, K. (2010). *Hybrid Methods fo rCoreference Resolution in Swedish (Doktorsavhandling).* Stockholm: Stockholms universitet.

Nivre, J. (2005). *Dependency grammar and dependency parsing. Technical Report.* Växjö University.

Patel, R., & Davidsson, B. (1994). *Forskningsmetodikens grunder - Att planera, genomföra och rapportera en undersökning.* Lund: Studentlitteratur.

Peirce, C. (Utgiven på svenska 1990). *Pragmatism och kosmologi.* Daidalos.

Popper, K. (1935). *Logik der Forschung.*

PowerSet. (n.d.). Retrieved 12 01, 2009, from http://www.powerset.com/

Ribeiro-Neto, R. A.-Y. (1999). Modern Information Retrieval. Boston, USA: Addison-Wesley Longman Publishing Co.

Ritz, J., Dipper, S., & Götze, M. (2008). Annotation of Information Structure: An Evaluation Across Different Types of Texts. *Proceedings of the 6th LREC-2008 Conference.* Marrakech.

Rogers, Y., Sharp, H., & Preece, J. (2013). *Interaction Design - Beyond Human-Compuer Interaction.* Wiley.

Rosell, M. (2005). *Automatisk synonymvariering av text.* Kursrapport, Språkgranskningsverktyg, KTH, Stockholm.

Rus, V., & Graesser, A. (2009). *The Question Generation Shared Task and Evaluation Challenge.* Memphis.

Sneiders, E. (2002). *Automated Question Answering: Template-Based Approach (Doktorsavhandling).* Stockholm: Stockholms universitet/KTH.

Stenmark, D. (2002). *Stenmark, D. (2002). Designing the new intranet (Doktorsavhandlilng), Gothenburg Studies in Informatics, report 21.* Göteborg: Göteborgs universitet.

Svenska Wikipedia. (n.d.). Retrieved 12 01, 2009, from http://sv.wikipedia.org

This is Watson (special issue). (2012, Maj-juni). IBM Journal of Research and Development.

Wilhelmsson, K. (2008). Automatic Variation of Swedish Text by Syntactic Fronting. *Workshop on NLP for Reading and Writing - Resources, Algorithms and Tools in conjunction with the SLTC Conference.* Stockholm: Webb-publikation: http://spraakbanken.gu.se/personal/sofie/SLTC_2008/.

Wilhelmsson, K. (2008). Heuristic Schema Parsing of Swedish Text. *SLTC 2008.* Stockholm.

Wilhelmsson, K. (2010). *Heuristisk analys med Diderichsens satsschema - tillämpningar för svensk text (doktorsavhandling).* Göteborg: Göteborgs universitet.

Wilhelmsson, K. (2011). Automatic Question Generation from Swedish Documents as a Tool for Information Extraction. *The 18th Nordic Conference of Computational Linguistics, NODALIDA.* Riga, Lettland.

Wilhelmsson, K. (2012). *Adverbialkarakteristik för praktisk informationsextraktion.* Göteborg: Adverbialkarakteristik för praktisk informationsextraktion i svensk text, GUISS.

Wilhemsson, K. (2012). Automatic question generation for Swedish: The current state. *Conference: The Workshop NLP for Computer-assisted Language Learning, Linköping Electronic Conference Proceedings, No. 80, SLTC 2012, Lund, Volume: No. 80.* Lund.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13 (2)*, pp. 260-269.