Lemma, lexem eller mittemellan?
Ontologisk ångest i den digitala domänen

Lars Borin

Ett digitalt dilemma

Hela Språkbanken ska skattskrivas. De omfattande text- och lexikonresurserna i Språkbanken <http://spraakbanken.gu.se> behöver standardiseras som en del av en upprustning av Språkbankens infrastruktur. Standardiseringen gäller både resursernas form och deras innehåll, men det är det senare som står i fokus här. Bland annat ska alla texter förses med lingvistisk uppmärkning eller *annotationer*: ordklass, böjningsform, vilken lexikonenhet som en viss textenhet skall hänföras till m.m.

De standardiserade annotationerna bör ha vissa egenskaper (här uppräknade utan inbördes rangordning). Det är således önskvärt att de är *formella* (i betydelsen axiomatiserade, för bl.a. inferens; se nedan), *lingvistiskt välmotiverade*, *deskriptiva och huvudsakligen empiriskt grundade*, *syftesanpassade*, *maskininlärbara* samt *uniformistiska* (dvs. de ska idealt gälla för alla stadier av alla språk).

Här behandlar jag enbart den lexikaliska informationen och mer specifikt frågan om vilken grundenhet man ska använda sig av i ett lexikon för språkteknologitillämpningar. För en sådan användning är det rent allmänt en klar fördel ifall man kan använda sig av en grundenhet som är konsekvent formellt definierad. Detta ska *inte* förstås som att definitionen uteslutande skulle få referera till språklig form och inte till språkligt innehåll, utan här avses enbart själva definitionens egenskaper; både språklig form och språkligt innehåll kan i princip karakteriseras i strikt formella termer.

Konkreta användningar i språkteknologiska system måste naturligtvis dock ofta ta hänsyn till en befintlig beskrivningstradition, speciellt om systemens användare direkt ska utsättas för de lingvistiska klassificeringar som systemen använder internt, såsom är fallet i t.ex. korpussökverktyg. Traditionen inom datalingvistik och språkteknologi liksom för övrigt inom stora delar av den konventionella lexikografin är att an-

vända sig antingen av grundformer (eventuellt plus ordklass; så görs t.ex. i SUC-korpusen och i många engelska lexikon) eller av semantiska grundenheter (i Wordnet använder man en traditionell kombination av innehåll och form som grundenhet men grupperar dessutom synonyma sådana enheter i något som man kallar *synset*).

I vår närmiljö finns ett förslag till grundenhet som det ligger nära till hands att överväga för användning i Språkbankens språkteknologiskt inriktade lexikon, nämligen *lemmat* i Språkdatas lemma-lexemmodell. Modellen tycks nämligen utlova en relativt kort väg till en formalisering av proceduren att bestämma ett språks grundläggande lexikonenheter.

Lemma-lexemmodellen maximalt mekaniserad

Lemma-lexemmodellen finns beskriven i ett antal arbeten. Mer utförliga beskrivningar ges av Allén (1967, 1970b, 1970c) och Burnage (1990), mer översiktliga av Allén (1970a, 1971, 1981), Berg (1978), Malmgren (1994) och Svensén (”det formella synsättet” 1987:197f.; ”den formellt-grammatiska metoden” 2004:122f.). Av de beskrivningar som rör svensk lexikografi är på det hela taget den tidigaste beskrivningen (Allén 1967) också den utförligaste; de efterföljande tillför i stort sett ingen substans till modellen, däremot ger de fler exempel och konkreta tillämpningsfall eller har ett uttalat pedagogiskt syfte (som Svensén 1987, 2004 eller den föredömligt klara framställningen i Malmgren 1994). Jag kommer i första hand att använda mig av Allén (1967) (rena sidangivelser inom parentes nedan hänvisar till detta arbete) som källa i mitt försök till algoritmisk rekonstruktion av lemma-lexemmodellen. Den ursprungliga framställningen är inte formell i den mening som man eftersträvar inom språkteknologin, där idealet är en detaljnivå som direkt ska kunna omvandlas i ett datorprogram (se ovan). Därför utgör många av detaljerna nedan närmast en tolkning av beskrivningen i Allén (1967) och andra av de ovan nämnda arbetena.

Endast formella språkkriterier – dvs. hämtade från språkets formsida, inte dess innehållssida – ska ligga till grund för bestämning av lemma-tillhörighet (s. 43), som omfattar ett antal steg:

(1) Generera en mängd av (konventionella) grundformer för lexikonord (enligt en viss uppfattning om vad som kan utgöra ett lexikonord).
(2) Utgående från denna mängd (1), definiera en tvåställig relation, en mängd av par (grundform, ordklass), där grund-

formernas ordklass(er) bestäms enligt en viss ordklassuppsättning och vissa kriterier för ordklasstilldelning (s. 46–55). Den nya mängden kommer att ha minst lika många element som utgångsmängden (en grundform kan tillhöra mer än en ordklass, t.ex. *sticka vb/nn*).

(3) Utgående från denna mängd (2), definiera en treställig relation, en mängd av lemmakandidater, trippler (grundform, ordklass, böjningsmönster) med användning av en viss uppsättning böjningsmönster. Den nya mängden kommer att ha minst lika många element som utgångsmängden (ett par av grundform plus ordklass kan motsvara mer än ett böjningsmönster, t.ex. *sticka vb stack/-de*).

(4) Fastställ en mängd av lemman genom parvis jämförelse av alla element i lemmakandidatmängden (3) med alla andra element i samma mängd, enligt en fastställd procedur (s. 55–62, särskilt figur 6, s. 60). Den nya (slutgiltiga) mängden kommer att ha högst lika många element som lemmakandidatmängden (vid det här laget kan kandidater bara slås samman, inte delas upp ytterligare).

(5) Definiera lexemen som de olika distinkta betydelser som ett lemma uttrycker. Lexemen är alltså hierarkiskt underordnade lemmana.

Ovanstående beskrivning är lätt att destillera fram ur källorna. Besvärligare är det att tolka de delar av proceduren som närmast framställs som axiomatiska. Ett behändigt sätt att beskriva aspekter av verkligheten formellt är ju att använda sig av en *axiomatisering*, ett logiskt system där ”alla använda begrepp explicit definieras med hjälp av ett antal på förhand angivna grundbegrepp [...] och alla påståenden (teorem) inom disciplinen härleds som logiska konsekvenser från ett antal på förhand angivna axiom [grundsatser som inte själva är föremål för bevis men som tjänar som utgångspunkt för bevis av andra satser]” (*Nationalencyklopedin*: s.v. *axiom*). Det axiomatiska systemet sett som modell av (eller teori om) ett stycke verklighet måste förses med en empirisk tolkning, som talar om hur systemets grundbegrepp, axiom, teorem och härledningsregler ska relateras till empiriska observationer av verkligheten (Dyvik 1980). Det ligger nära till hands att betrakta lemma-lexem-modellen som ett förslag till en axiomatisering av den lingvistiska lexikaliska domänen. Även om denna tolkning är naturlig, får man konstatera att den faktiska framställningen i källorna inte är deduktiv som man skulle ha förväntat sig, utan närmast abduktiv: Vad ett lemma är

illustreras med exempel, och lagbundenheter antyds likaledes med exempel.

Axiomatiska aspekter av lemma-lexemmodellen

Modellen innehåller flera delar som måste betraktas som grundbegrepp eller axiom, alltså tagna för givna och inte härledda inom modellen, eller åtminstone inte förklarade i källorna. Dit hör hur den första mängden av grundformer genereras. Dit hör kriterierna för vilka ordklasser som ska antas. Dit hör böjningsmönstren, som är hämtade från *Illustrerad svensk ordbok*.[1] Dit hör också vad som ska uppfattas som ett lexikonord.[2] Dit hör dock framför allt de grundläggande kriterierna i den jämförelse som ytterst ska avgöra om man har att göra med ett eller flera lemman, nämligen dels vilka språkliga formkarakteristika som man ska ta hänsyn till när man jämför lemmakandidater, dels ”det i lingvistisk analys övliga kommutationsprovet” (s. 55). För språkteknologisk användbarhet skulle det sistnämnda provet dock behöva formaliseras så det blir mekaniskt avgörbart när ”klassen av formernas kontexter” (s. 55) är densamma för två former. Frågan är om valet av kommutationsprovet som grundläggande test och själva hjärtat i lemma-lexemmodellen – som *alla* lemmakandidater ska genomgå – tillför modellen något ifråga om stringens. Jag skulle hävda att så inte är fallet, eftersom provet i praktiken är ogenomförbart – inte minst därför att antalet potentiella kontexter är oändligt – och således kommer att reduceras till ett intuitionsomdöme om betydelselikhet. Sålunda finns det såvitt jag kan se inget i framställningen i Allén (1967) som *formellt* förhindrar att t.ex. serierna *syssling -en -ar* och *tremänning -en -ar* förs ihop till ett lemma (med fri stamvariation). Lemma-lexemmodellen visar sig i grund och botten vara en lexikalisk modell baserad på betydelse, med alla de gränsdragningsproblem detta medför.

I det nederländska CELEX-projektet har man använt sig av en modell som mycket liknar Språkdatas lemma-lexemmodell, men som avviker

---

[1] Man kan därvid notera att det är oklart om suppletiv böjning tillåts; i flera källor sägs att den inte ska vara tillåten, men en del exempel i Allén (1967) antyder att även suppletiva flexionsserier tillåts.

[2] I Språkbanken ser vi av flera skäl ett behov av att kunna hantera flerordsenheter av olika slag på samma sätt som enkla ord, inklusive tilldela dem grundformer. Beskrivningarna av lemma-lexemmodellen är här inte helt entydiga. Allén (1975) verkar inte räkna med flerordslemman, medan i andra källor den möjligheten verkar finnas (t.ex. Anon. 1995).

på två viktiga punkter, en substantiell och en metodologisk: Dels används en annan uppsättning språkliga formkarakteristika som lemmaskiljande kriterium, varom mer nedan, dels använder man betydelseskillnad explicit axiomatiskt; det antas helt enkelt att det finns ett sätt att avgöra om betydelseskillnad föreligger (Burnage 1990). Metodologiskt innebär detta att man aldrig gör rena formjämförelser, utan alltid jämför språklig form och betydelse parallellt. Har man anammat Saussures tanke om det språkliga tecknet som en oupplöslig förening av form och innehåll, så är detta inte ägnat att förvåna; man kan inte förvänta sig att i språket finna ren form bortom ordens fonetiska/grafiska gestalt. Även ett eventuellt studium av ren betydelse faller utanför grammatikens och lexikologins domvärjo.

Om man ersätter kommutationsprovet med ett ”orakelomdöme” om betydelseskillnad à la CELEX-modellen så kan man också eliminera ett lingvistiskt omotiverat drag i lemma-lexemmodellen, nämligen att den inte erkänner paradigm som helt och hållet ingår i andra paradigm, t.ex. substantiv utan pluralformer eller verb utan perfektparticip. Det framgår inte alldeles klart av framställningen (”Flexionsserier baserade på frånvaro av viss form får avvisas.” s. 59) om det handlar om ett generellt förbud mot sådana paradigm eller ett mer begränsat förbud mot att ansätta mer än ett lemma i fall som *golf -en* (’en sorts spel’) och *golf -en -er* (’havsvik’), eller *sprita vb (tr) -de* (’lossa [frö] från fäste’) och *sprita vb (it) -de* (’dricka sprit’).[3] Förbudet följer direkt av beslutet att använda kommutationsprovet, eftersom – som Allén mycket riktigt konstaterar – man generellt inte kan använda provet för att jämföra närvaron av ett element med dess frånvaro.

Ontologiska överväganden

Även om lemma-lexemmodellen inte är formaliserad till den grad man hade önskat, kan det ju mycket väl vara så att den fångar någon viktig lingvistisk generalisering eller annars resulterar i en beskrivning med önskvärda egenskaper och därför skulle väljas framför en annan, likaledes ickeformell modell.

En önskvärd egenskap skulle vara att den ger som resultat en beskriv-

---

[3] Både *Nationalencyklopedins ordbok* och *Svenska Akademiens ordlista* (trettonde upplagan, 2006) anger två uppslagsord *golf* men ett uppslagsord *sprita*. Intressant nog brukar inte lingvistiska teoretiska verk om morfologi alls beröra paradigm helt inneslutna i andra paradigm (t.ex. Haspelmath 2002), något som vi ändå empiriskt vet förekommer. Här finns uppenbarligen utrymme för mer forskning.

ning som är strikt nivåindelad, något som ofta sägs om modellen. Vi skulle få en beskrivning som reser vattentäta skott mellan språklig form (lemmat) och språkligt innehåll (lexemet), vilket dessutom skulle fånga en distinktion med lång tradition inom lingvistiken. Det skulle praktiskt innebära att man aldrig för ett lexem skulle behöva ange någon information som hänvisar till språklig form; allt som behövs i den vägen skulle lexemet ha ärvt från sitt överordnade lemma. Det är givetvis så att både form och innehåll måste användas när enheterna fastställs; det just sagda handlar enbart om hur man sedan beskriver dem. Men då fordras en klar och uttömmande bild av vilka språkdrag som ger till resultat en beskrivning med de önskade egenskaperna. För lemma-lexemmodellen anges dessa vara fonematisk form, grafematisk form, ordklass och böjningsklass (s. 47; ett femte kriterium som anges där, lemmatisk klass, är i grunden ett innehållskriterium). I CELEX-modellen används fler formkriterier, bl.a. morfologisk struktur, som används för att skilja eng. *rubber nn* (’sb/sth that rubs’ [rub+er]) från *rubber nn* (’an elastic substance’ [enmorfemigt]).

Varken Allén (1967) eller Burnage (1990) motiverar de kriterier man valt, trots att det inte är svårt att komma på andra aspekter av språklig form som skulle kunna vara relevanta (och förstås även många som inte alls är relevanta). För svenskans vidkommande har vi t.ex. sammansättningsform (som faktiskt är paradigmdefinierande för Hellberg 1978), (syntaktisk) valens hos verb och (in)dividualitet hos substantiv. En intressant fråga är i vilken grad det är möjligt att ställa upp en uttömmande lista av relevanta kriterier.

Exempelvis morfologisk struktur skulle kunna vara relevant även för svenska, kanske framför allt för sammansättningar. Om man gör det rimliga antagandet att sammansättningsled i formhänseende hör hemma på lemmanivån snarare än på lexemnivån, så skulle sammansättningar som *valbevakning*, *maskbärare* m.fl. generera två lemman, eftersom de i så fall har (åtminstone) två olika morfologiska analyser vardera.[4]

Eventuellt är det så att lemma-lexemmodellen faktiskt lämpar sig bäst för den verksamhet i vilken den ursprungligen kom till, nämligen utarbetande av lexikon för mänskligt bruk. I ett sådant lexikon benämner lemmat en lexikonartikel med ett rikt informationsinnehåll. För språk-

---

[4] Allén (1980) väljer en annan lösning och inför termen *formellt morfem* för bl.a. sammansättningsled som *val-*. Vare sig det formella morfemet eller Hellbergs (1978) *tekniska stam* kan dock såvitt jag kan se motiveras som lingvistiska enheter. Detta eftersom båda utgör ren form utan beledsagande betydelse; lemmat däremot framstår som en respektabel lingvistisk enhet just därför att dess definierande kriterier refererar till både språklig form och språkligt innehåll.

teknologisystem behöver man eventuellt en mer ”heltäckande” lemma-definition än modellen tillhandahåller. Det kan uppstå – och uppstår – behov av att kunna annotera alla slags textord med lemmatillhörighet, inklusive produktivt bildade sammansättningar och avledningar (t.ex. adverb från adjektiv och eventuellt particip från verb). Det betyder att över tid kommer dessa att helt dominera i lemmabeståndet, medan många av dem överhuvudtaget inte anses höra hemma i ett vanligt lexikon. I Språkbanken behövs framför allt en identifierare, ett ”person-nummer”, med två funktioner: (1) som index i en lexikalisk databas och (2) för att identifiera (vad man vill betrakta som) samma enhet över många olika digitala språkliga resurser. Det är mot den bakgrunden kanske inte så förvånande att man i lemma-lexemmodellen inte har tagit ställning till en del knepiga frågor rörande sammansättningar (och avledningar), men att man i en språkteknologikontext är piskad att göra det.

Lemma, lexem eller mittemellan?

Vart har denna lilla begreppsutredning fört oss? Jag tror att lemma-lexemmodellen på det stora hela pekar i rätt riktning och på ett fruktbart sätt vidareutvecklar den metodologiska distinktionen mellan språklig form och språklig betydelse. Det är säkert så att dess anspråk på större formaliserbarhet och därmed större formalisering har lett till en välbehövlig skärpning i svensk lexikografisk praxis och därmed varit till stor nytta för det lexikografiska arbetet vid Språkdata, som ju har avkastat flera stora moderna lexikon för svenska.

Att kunna behandla språkliga enheters form som konceptuellt separerad från deras innehåll har man stor nytta av i många sammanhang. Däremot kommer man aldrig ifrån att en av den moderna lingvistikens hörnstenar är axiomet att form och innehåll är oupplösligt förenade i det språkliga tecknet. Därför är det en chimär att tro att man ska kunna definiera någon genuint språklig enhet åberopande enbart den ena av det språkliga tecknets två aspekter.

Paradoxalt nog ger detta oss faktiskt mer frihet vid utformningen av de lexikaliska referenssystemen i Språkbanken. Det kanske rentav kommer att visa sig ändamålsenligast att låta såväl form – någon sorts ”lemma” – som innehåll – någon sorts ”lexem” – vara full- och likvärdiga medborgare i dem, fast alltid logiskt förbundna med varandra. Det faktum att det ofta är svårt att avgränsa betydelser är i princip inte mer förödande för ett sådant företag än det är för förfärdigandet av en ordbok som ska innehålla ordförklaringar.

Litteratur

Allén, Sture 1967. Studier över nusvenskans vokabulärsystem. Opubl. rapport. Institutionen för nordiska språk, Göteborgs universitet.

Allén, Sture 1970a. En undersökning av nusvenskans vokabulärsystem. I: Sture Allén & Jan Thavenius (utg.), *Språklig databehandling. Datamaskinen i språk- och litteraturforskning*. Lund: Studentlitteratur. S. 31–59.

Allén, Sture 1970b. Vocabulary data processing. I: Hreinn Benediktsson (ed.), *The Nordic Languages and Modern Linguistics*. Reykjavík: Vísindafélag Íslendinga. S. 235–261.

Allén, Sture 1970c. Inledning. I: Sture Allén, *Nusvensk frekvensordbok baserad på tidningstext. 1: Graford. Homografkomponenter*. Stockholm: Almqvist & Wiksell. S. XIII–XXX.

Allén, Sture 1971. Inledning. I: Sture Allén, *Nusvensk frekvensordbok baserad på tidningstext. 2: Lemman*. Stockholm: Almqvist & Wiksell. S. XIII–XXVII.

Allén, Sture 1975. Inledning. I: Sture Allén et al., *Nusvensk frekvensordbok baserad på tidningstext. 3: Ordförbindelser*. Stockholm: Almqvist & Wiksell. S. XIV–XXX.

Allén, Sture 1980. Inledning. I: Sture Allén, Sture Berg, Jerker Järborg, Jonas Löfström, Bo Ralph & Christian Sjögreen, *Nusvensk frekvensordbok baserad på tidningstext. 4: Ordled. Betydelser*. Stockholm: Almqvist & Wiksell. S. XV–XXXII

Allén, Sture 1981. The lemma-lexeme model of the Swedish Lexical Data Base. I: Burghard B. Rieger (ed.), *Empirical semantics II*. Bochum: Brockmeyer. S. 376–387.

Anon. 1995. Inledning. I: *Nationalencyklopedins ordbok. Utarbetad vid Språkdata, Göteborgs universitet* 1. Höganäs: Bra Böcker. n. pag.

Berg, Sture 1978. *Olika lika ord. Svenskt homograflexikon*. Stockholm: Almqvist & Wiksell.

Burnage, Gavin 1990. CELEX: A guide for users. CELEX – Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen, <www.ru.nl/celex/subsecs/celex_intro.pdf>.

Dyvik, Helge 1980. *Grammatikk og empiri. En syntaktisk modell og dens forutsetninger*. Universitetet i Bergen, Institutt for fonetikk og lingvistikk, Skriftserie, Nr. 24, Serie B.

Haspelmath, Martin 2002. *Understanding morphology*. London: Arnold.

Hellberg, Staffan 1978. *The morphology of present-day Swedish*. Stockholm: Almqvist & Wiksell.

Malmgren, Sven-Göran 1994. *Svensk lexikologi. Ord, ordbildning, ordböcker och orddatabaser*. Lund: Studentlitteratur.

Svensén, Bo 1987. *Handbok i lexikografi*. Stockholm: Esselte Studium & Tekniska nomenklaturcentralen.

Svensén, Bo 2004. *Handbok i lexikografi*. 2 omarbetade och utökade upplagan. Stockholm: Norstedts Akademiska Förlag.