SPRÅKDATA
Rolf Gavare
1979-09-10

AUTOMATISK LEMMATISERING UTAN STAMLEXIKON

Några synpunkter tio år efteråt

En tillbakablick

För precis tio år sedan, hösten 1969, genomfördes det första stora lemmatiseringsarbetet vid Språkdata. Det var den bearbetning som kom att bilda grundmaterialet till Nusvensk frekvensordbok 2. Två år senare gjordes en lemmatisering av Svenska psalmboken, med oförändrad programvara men med lite andra förutsättningar, vilket jag skall återkomma till. Trots att såväl datortekniken som de datalingvistiska metoderna utvecklats starkt under de år som gått, har erfarenheterna från det nämnda projektet fortfarande stor aktualitet. Jag skall först belysa några av de speciella omständigheter som gällde vid de båda nämnda körningarna, för att sedan skissera en modern implementering av interaktiv lemmatisering, som en vidareutveckling av den gamla modellen.

Grunddragen i lemmatiseringsmodellen framgår av Staffan Hellbergs bifogade artikel Computerized Lemmatization without the Use of a Dictionary. Sammanfattningsvis krävs enligt denna modell att de komponenter som skall ingå i lemmat uppfyller de tre huvudkriterierna: (1) alla enheterna skall ha en identisk stam, (2) ändelserna skall tillhöra samma paradigm och (3) eventuella ordklass/lemmabeteckningar skall vara lika och dessutom passa till samtliga komponenters ändelser. Några speciella förutsättningar som gällde det aktuella projektet bör hållas i minnet. Algoritmen var utformad för att användas på ett alfabetiskt sorterat, tidigare homografseparerat material. (Att det var alfabetiskt ordnat betyder förstås inte att de enheter som skall grupperas samman behövde stå intill varandra.) Det material som redovisas i Nusvensk frekvensordbok omfattade 1 miljon löpande ord, vilket gav 103 000 olika graford och 112 000 homografkomponenter (cirka 30 % av de olika graforden var homografa). Nära 97 % av homografkomponenterna placerades automatiskt i rätt lemma. Av de rätt avgränsade lemmana fick 85 % också rätt rubrik med uppslagsform, ordklass- och lemmabeteckning. Svenska psalmbokens 8500 olika graford (av c. 110 000

Automatisk lemmatisering utan stamlexikon
Rolf Gavare
Proceedings of NODALIDA 1979, pages 123-132

löpande ord) var inte homografseparerade och gav därför ett sämre resultat – omkring 80 % av de oseparerade orden hamnade i rätt lemma (d.v.s. rätt för åtminstone någon homografkomponent av det grafiska ordet). Ytterligare skäl till att psalmbokslemmatiseringen gav sämre utfall är att paradigmtabellerna var utformade med tanke på sannolikheter för uppträdande i modern svenska och framför allt, ju större ordmaterialet är, ju fler olika böjningsformer är belagda inom varje lemma, vilket i sin tur ger en större säkerhet vid såväl sammanföringen av böjningsformer som etableringen av en lemmarubrik. Algoritmen är inte utformad för lemmatisering av enstaka ord.

Bland de tekniska förutsättningarna märks särskilt att vi vid den aktuella tidpunkten saknade direktaccessminne (skivminne). I stället fick magnetbandstationer utnyttjas för lagring av arbetsfiler, vilket också i viss mån kom att återspeglas i programlogiken. Programutrymmet var också begränsat. Någon möjlighet till interaktion via terminaler fanns inte. Trots de praktiska begränsningarna får man ändå anse att den automatiska lemmatiseringen lyckades väl.

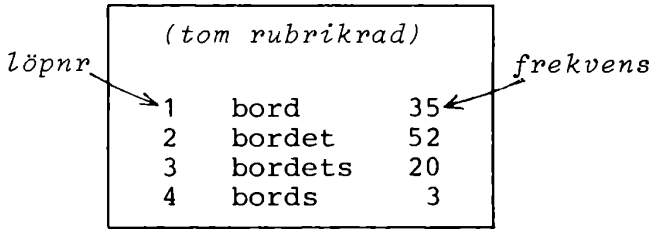
Dagsläget

Jag skall nu övergå till att skissera huvuddragen av hur en lemmatisering (inklusive homografseparering), byggd på dessa principer, skulle kunna gå till idag.

1. Antag att vi har ett stort, obearbetat textmaterial (på grafordsnivå). En första åtgärd blir att ta fram en komplett konkordans i radskrivarutskrift (eller på mikrokort). Om möjligt skall konkordansen också vara tillgänglig i datorn för sökning direkt från terminalen. Både versaler och gemena tecken skall återges i utskriften för att underlätta bedömningen av egennamn, initialförkortningar och versal i meningsbörjan. Konkordansen skall vara ordinärt alfabetiskt sorterad på de alfabetiska tecknen i stickordet och högerkontexten. Beläggen av samma (normaliserade) graford förses med en löpnumrering. En sådan konkordans har ett mycket stort värde även bortsett från den här avsedda lemmatiseringen/homografsepareringen.

2. En ändelse-/paradigmlista upprättas i enlighet med Staffan Hellbergs alternativa metod. Något stamlexikon skall inte användas.

3. Lemmatiseringen (utan föregående homografseparering) försiggår on-line vid en textskärmsterminal. Detta förfarande minimerar omfattningen av felspridning vid lemmatiseringen. De sublemman (graford) som programmet anser skall bilda det aktuella lemmat visas på skärmen, exempelvis på det här sättet:



Lemmatiseringsalgoritmen bygger i huvudsak på Staffan Hellbergs alternativa modell. Eventuellt kan denna kompletteras för att ge ett 'intelligent' förslag till homografseparering automatiskt med hjälp av särskilt homograflexikon och viss kontextkontroll.

4. Med enkla kommandon skall det aktuella lemmat kunna modifieras:
   a. sublemman skall kunna tas bort ur lemmat (och behandlas senare),
   b. sublemman skall kunna flyttas till ett tidigare lemma (rubriken anges och lemmat ifråga visas på skärmen för kontroll),
   c. sublemman skall kunna separeras (vissa belägg (vars beläggställen hämtas ur konkordansen) bildar en ny sublemmaform, som sedan eventuellt flyttas ur lemmat),
   d. sublemman skall kunna märkas för intern homografi och ev. polysemi,
   e. tidigare behandlade lemman skall när som helst kunna inspekteras, modifieras och skrivas ut.

5. Då lemmats sublemmauppsättning godkänts föreslår programmet en lemmarubrik med ordklass- och lemmabeteckning.

6. Efter kontroll och eventuell ändring av lemmarubriken presenteras nästa lemma på skärmen (enligt punkt 3).

7. Helst bör möjlighet finnas att dynamiskt uppdatera paradigm/ändelseregistret under körningens gång.

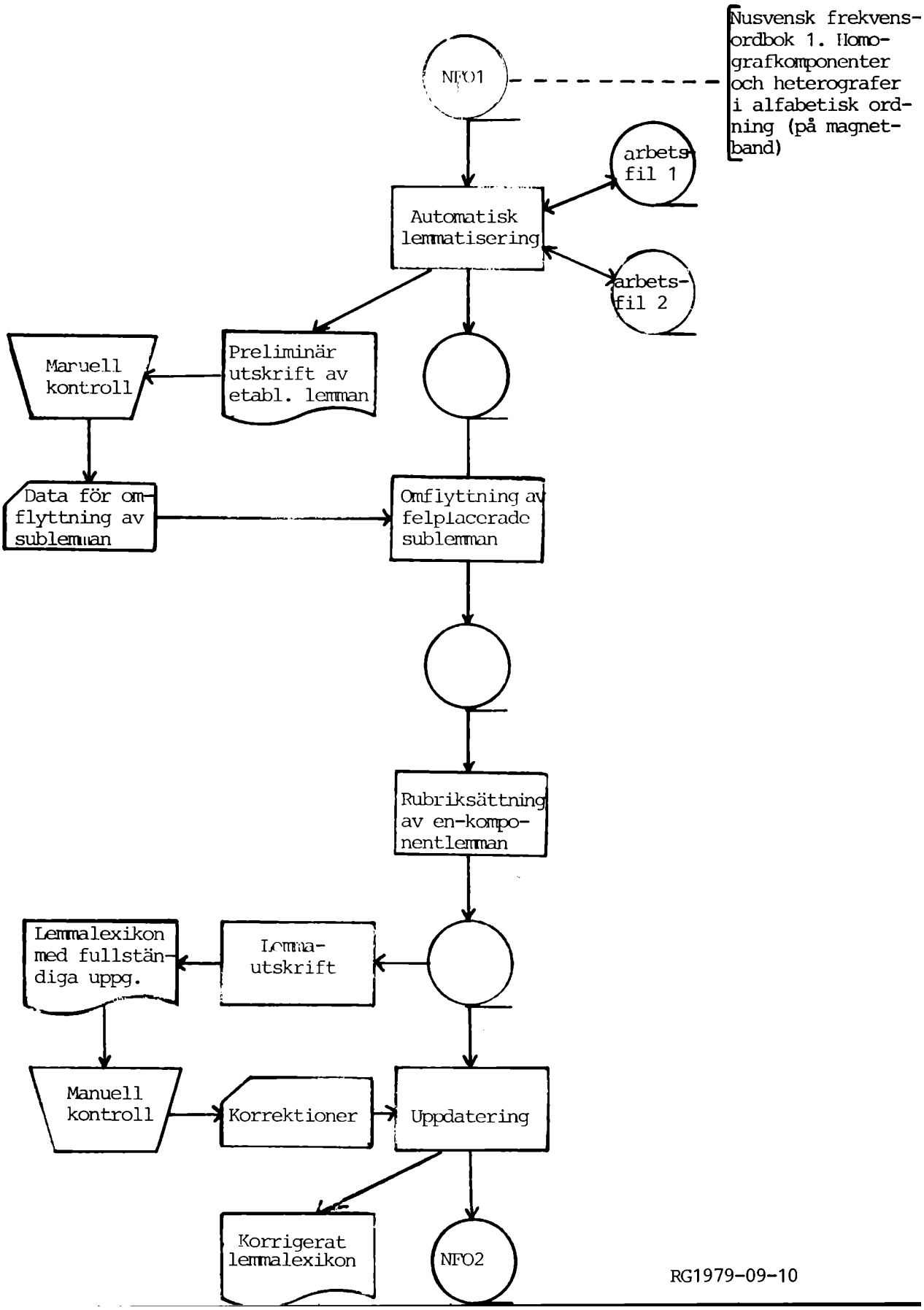
Det skisserade förslaget är utformat med tanke på krav på språkvetenskaplig noggrannhet. Helautomatisk lemmatisering med lägre krav på korrekthet kan förstås också i vissa fall vara önskvärd.

Att etablera grundformen till enstaka ordformer ställer andra krav på modellen. För att få en god säkerhet i detta fall behövs förmodligen ett morfem- eller stamlexikon, exempelvis av den typ som användes inom projektet Algoritmisk textanalys, se vidare Staffan Hellbergs bok The Morphology of Present-Day Swedish (Data linguistica 13, Almqvist & Wiksell International, Stockholm 1978).

51530	FOTOGENLAMPA-	NN -N			
1	FOTOGENLAMPA		0		
2	FOTOGENLAMPAN		0		
51540	FOTOGRAF-	NN -EN			
1	FOTOGRAF		0	*	
2	FOTOGRAFEN		0	EN	
3	FOTOGRAFENS		0	EN	
4	FOTOGRAFER		0		
5	FOTOGRAFERNA		0		
6	FOTOGRAFERNAS		0		
51550	FOTOGRAFER-A	VB -AD			
1	FOTOGRAFERA	VB -AD INP			
2	FOTOGRAFERA	VB -AD INF			
3	FOTOGRAFERAD		A		
4	FOTOGRAFERADE	VB -AD PRT			
5	FOTOGRAFERADE	VB -AD PTP			
6	FOTOGRAFERADES	VB -AD PRT			
7	FOTOGRAFERANDE	VB -AD			
8	FOTOGRAFERAR		A		
9	FOTOGRAFERAS	VB -AD INF			
10	FOTOGRAFERAT	VB -AD SUM			
51560	FOTOGRAFERING-	NN -EN			
1	FOTOGRAFERING		0	*	
2	FOTOGRAFERINGEN		0	EN	
51570	**********-				
1	FOTOGRAFERINGSFÖRBUDET				
51580	FOTOGRAFI-	NN -ET			
1	FOTOGRAFI		0		
2	FOTOGRAFIER		ER	0	UM
3	FOTOGRAFIERNA		0	UM	
4	FOTOGRAFIET		0	UM	
51590	FOTOGRAFISK-	AV -T			
1	FOTOGRAFISK				
2	FOTOGRAFISKA	AV -T NEU	0		
3	FOTOGRAFISKA	AV -T PLU			
51600	**********-	AB			
1	FOTOGRAFISKT	AB			
51610	**********-				
1	FOTOHISTORIA				

NUSVENSK FREKVENSORDBOK 2

Lemmatiseringsfasen (1969/70)



Computerized Lemmatization without the Use of a Dictionary: A Case Study from Swedish Lexicology

Staffan Hellberg

Lemmatization, i.e., the bringing together of the inflectional forms (and variant forms) of a word under one heading, is one of the problems when making a frequency dictionary out of a large text corpus with the aid of a computer. Attempts have generally gone in the direction of confronting the material with an ordinary dictionary, presupposing that this dictionary would have an entry for practically every form in the corpus. This may be true for some texts, e.g., the classical ones, but it is definitely not true for a newspaper text corpus in a language like Swedish, which not only shows brand-new loan-words but is abundant in compounds of the more or less casual sort that will never appear in ordinary dictionaries. So the task we undertook in 1969-70 at the Research Group for Modern Swedish, University of Göteborg, was to lemmatize automatically about 112,000 different word forms without direct access to any existing dictionary. Homographs had been previously separated with the aid of a KWIC-index (the original number of different graphic words was about 103,000), a fact which meant that about one-third of the forms had been assigned grammatical information (word class and, roughly, gender or conjugation).[1]

As Swedish contains no inflectional prefixes, the procedure can operate with an alphabetically sorted version of the material. The computer passes through that version, successively grouping the forms into lemmas and printing them out, so that the whole lemmatization can be checked manually afterwards.

The program tests the first form in the projected lemma against the ones following alphabetically, one at a time, providing they have not already been included in a previously finished lemma. As soon as a form appears which is not identical with the first form as far as the stem of the latter goes, the testing is stopped and the lemma finished. The form that heads the remainder of forms alphabetically is then chosen as the first form of the next lemma, and the procedure is repeated.

[1] Allén, *Nusvensk frekvensordbok/Frequency dictionary of present-day Swedish 1*, Stockholm: Almqvist and Wiksell, 1970, presents the material on the level where homographs had been separated but no lemmatization done. The lemmatized version was published in 1971 as Allén, *Nusvensk frekvensordbok/Frequency dictionary of present-day Swedish* 2. A more detailed report of the lemmatization has been made (in Swedish) in Staffan Hellberg, *Automatisk lemmatisering*, 1971 (mimeographed). A general survey of the work at the Research Group will be found in Sture Allén, "Vocabulary data processing," *Proceedings of the International Conference of Nordic and General Linguistics, Reykjavik, 1969*, ed. Hreinn Benediktsson, Reykjavik 1970.

Staffan Hellberg is a member of the Research Group for Modern Swedish at the University of Göteborg.

The "stem" of a lemma thus had to be defined as the part of the word that was identical in all its inflectional forms. The remnants of the forms were called "endings." Obviously, these definitions don't altogether coincide with the usual linguistic ones: the word *titel,* 'title,' plural *titlar,* for instance, got the stem *tit-* and the endings *-el, -lar,* etc., though linguistically, the stem should rather be *titl-* and the plural ending *-ar.* An index was set up of those graphic sequences that might be endings in regular paradigms.[2] Lexical regularity proved not to be the same thing as grammatical regularity; for instance, an irregular noun occurring as the latter element in many compounds had to be taken account of. The word *man* 'man,' plural *män,* is as irregular as in English, but it appeared in over 150 compounds in the corpus, e.g., *adelsman,* 'nobleman,' plural *adelsmän,* and so a paradigm *-an, -än,* etc. was established. In all, 53 different paradigms were made the basis of the index, which contained 98 different endings. The figures give a somewhat exaggerated idea of the complexity of Swedish morphology, as one linguistic paradigm often had to be split into two or more paradigms here: compare *titel, titlar* (above) with the endings *-el, -lar,* etc. to *stol,* 'chair,' plural *stolar* with the endings *-ø, -ar,* etc.

For two forms to be brought into the same lemma, they were required to have an identical stem and compatible endings, i.e., such as could belong to the same paradigm. Whether the identity actually covered the whole stem was decided by checking whether the remnants of the forms were possible endings. So the index here served two purposes: to identify the latter parts of the forms as endings, and to give access to what was called the alpha-list, where for each ending the endings compatible with it were stored. But for the former procedure to function properly, it was necessary that every graph or graphic sequence Y which could not itself be an ending but which had a counterpart XY that was a possible ending appear in the index, where it was stored as a pseudo-ending with an empty alpha-row. An example is the final *-l* which didn't occur in any paradigm, while the sequence *-el* did (see *titel, titlar* above). In all, 14 pseudo-endings were required.

If the alpha test gave a negative result, it was repeated with the rightmost graph (roughly: letter) of the stem brought over to the endings, provided, of course, that these new endings were to be found in the index at all. But once a shorter stem had been recognized by a successful test of that kind, it was not allowed to be lengthened again as a result of a comparison with yet another form, because that would mean an obvious mixing of two paradigms.

The index served its third purpose when giving entry to the so-called beta-list, where the possible grammatical labels were given for each ending. The beta-list was consulted when one of the tested forms, or both, was a homograph and thus "marked" for grammatical category, and so a number of wrong lemmatizations could be prevented through the demand for grammatical compatibility. The beta-list was also used in the subprogram of automatic attributing of head forms and grammatical labels to all the lemmas, which will not be reported here.

The main course of the procedure is shown in the flow chart. Several improvements were suggested by our programmer, Rolf Gavare, who wrote the program in DATASAAB/ALGOL–GENIUS and DAC.

Some measures were taken to compress the lists. One of these made use of the structure of the Swedish inflectional system, where the ending *-s* plays a unique role. It always occupies the last position in the form, and it can be added to practically every form of nouns, adjectives, and verbs, having either a genitive or a passive function. If all those *s*-variants of the endings had been accounted for in the normal way, it would have meant nearly a doubling of the index and a considerable enlarging of the alpha-list.

[2] The bulk of the paradigms were taken from Björn Hammarberg, "Maskinell generering av böjningsformer och identifikation av ordklass," *Förhandlingar vid sammankomst för att dryfta frågor rörande svenskans beskrivning 3, Göteborg, 1965,* ed. Sture Allén, Göteborg, 1966 (mimeographed).

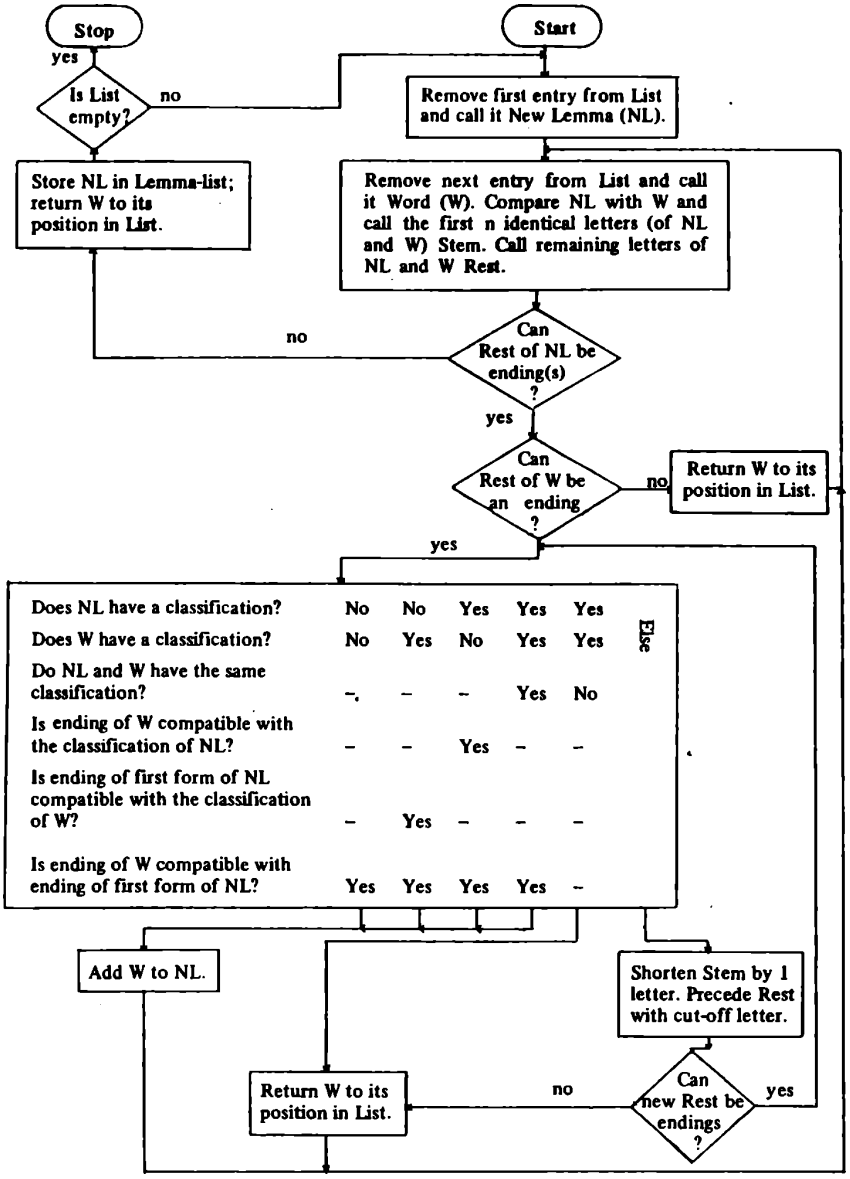


Figure 1

Instead, all forms ending with an *-s* were treated as if the *-s* wasn't there, except those where the *-s* belonged to the stem and which could be readily sorted out, as they were homographs internally with their own genitives and thus had a special "marking."

There were also quite a few *ad hoc* measures taken to obtain a better result, as several minor defects could be foreseen during the construction of the lists and by scrutinizing the result of test computations. Some of the measures simply meant omitting

an item from one of the lists, thereby replacing a number of wrong lemmatizations by a smaller number of missing correct ones. A measure of a different kind worth mentioning was the rearrangement of the alphabetically ordered material so that out of two homographs, one noun and one verb, the verb was placed before the noun. That saved a fair number of lemmas from going wrong.

The lemmatization yielded about 71,000 lemmas in all. The figure reveals that a large number of lemmas appeared in only one form. These lemmas did not cause any special troubles to the program, as a projected lemma could often be finished after its first form had been compared to—and shown too little similarity to—its nearest neighbor in the alphabetical order. A different subprogram had to be designed, though, for the attribution of head forms and grammatical labels (see above), as the beta-list gave no information in this case, where no boundary between stem and ending had been definitely established.

Though the whole corpus was treated in the manner now described, not all lemmas could, of course, be made to come out correctly from the computer. The program would have been hopelessly slow and complex if it had had to account for strong verbs, regular though they might be. There were also very rare paradigms that would have done more harm than good if they had been brought into the lists. In fact, the accomplished wrong lemmatizations are more notable than the missing correct ones. Not all clashes could be prevented by the above-mentioned *ad hoc* measures. And as the material also contained foreign words occurring in the newspaper corpus, there appeared a number of ridiculous lemmas, such as the one consisting of (English) *fair* and (French) *faire.*

The manual check of the computer output showed that 3.5 percent of the forms were in the wrong place and had to be moved. As this check was done with relative ease, the lemmatization program may well be said to have saved us from a considerable amount of dull routine work. Still, it could be asked whether the automatic procedure has actually been optimized. The number of wrong lemmatizations indicates that the alpha-list didn't have a sufficient discriminating function. This is actually natural for Swedish, where some sequences are very common as endings in different functions: the ending *-er,* for instance, occurs in 12 paradigms and is compatible with 29 other endings.

In closing, I will give a brief account of an alternative solution that I outlined after the computing of our material had been accomplished. In this solution, the ideas of alphabetical procedure and of an index of possible endings are taken over from the system used. But the alpha- and beta-lists are replaced by what could be called the gamma-list. That is, for each ending information is now given about which paradigms this ending can occur in, the paradigms having numbers from 1 to 53. For two forms to be brought together, it is now required that they have an identical stem and at least one paradigm number in common. If the common number or numbers are stored, a third form can be tested against them, and that means that any new tentative form will be tested against all the previously accepted forms in the lemma, which wasn't possible in the system used.

The beta-list is made superfluous by the grammatical labels being brought into the index and assigned paradigm numbers. So when two forms are tested, one of which is a homograph, it is required that at least one number occur three times in the gamma-list: with the two endings and with the grammatical label.

Most of the measures taken to improve the system used can be kept, as for instance the special treatment of forms ending in inflectional *-s.* Though the alternative solution hasn't been tested on the material, it seems fairly clear that it would have surpassed the one we chose. Of 12 different kinds of clashes that had been registered before the new system was developed, seven would have been avoided. What this would mean in figures is harder to guess. A reduction of the number of wrong lemmatizations by one-half is perhaps a somewhat too optimistic estimation.