# Changing the tokenization in Talbanken to SUC2.0

Bengt Dahlqvist & Beata Bandmann Megyesi
Department of Linguistics and Philology
Uppsala University, Uppsala, Sweden.

14 maj 2007

We describe our work on converting the tokenization of Talbanken using the same principles as the tokenization of the second version of the Stockholm Umeå Corpus (SUC2.0).

Abbreviations consisting of several tokens, as any other multiword expressions in Talbanken are split up into tokens, each represented on separate line, as shown in the examples below. The first token (second column) of the expression is annotated with the part-of-speech tag (third column) the expression belongs to, and the other tokens in the expression receive an ID tag.

```
P10101003004    BL.             ABZA    CA      003
P10101003005    A.              ID      CA      003

P10105010005    T.              ABZA    CA      009
P10105010006    EX.             ID      CA      009

P10107015017    FR.             PR      TAPR    014
P10107015018    O.              ID      TAPR    014
P10107015019    M.              ID      TAPR    014

P10106012007    I               PR      TBPR    011
P10106012008    SAMBAND         ID      TBPR    011
P10106012009    MED             ID      TBPR    011

P10109021002    *NÄR            PR      AAPR    020
P10109021003    DET             ID      AAPR    020
P10109021004    GÄLLER          ID      AAPR    020

P10120048015    VID             PR      AAPR    046
```

```
P10120048016     SIDAN                    ID      AAPR        046
P10120048017     AV                       ID      AAPR        046
```

To find possible abbreviations, we automatically extracted the tokens annotated with ID tags and the previous head token bearing the part-of-speech of the expression, and converted these into one single token where the included tokens were separated by "_". In this way, we extracted 724 multiword expressions in Talbanken which are listed in the end of this document. Then, we manually checked the list containing these expressions and extracted 29 abbreviations in total. The found abbreviations are shown below.

```
Bl._a.
Bl_a
Fr_o_m
T._ex.
bl._a.
bl_a
d._v._s.
d_v_s
e._d.
e_d
f._n.
f_n
fr._o._m.
m._fl.
m_fl
m._m.
m_m
o._s._v.
o_s_v
s._k.
s_k
t._ex
t._ex.
t_ex
t._h.
t._o._m.
t_o_m
t._v.
t_v
```

Lastly, on the basis of the abbreviation list, we concatenated the involved tokens into one according to SUC standard in terms of punctuation. Below,

the rules for re-writing the abbreviations in Talbanken into SUC standard is follows.

| Talbanken | SUC |
|---|---|
| Bl._a. | Bl.a. |
| Bl_a | Bl_a |
| Fr_o_m | Fr_o_m |
| T._ex. | T.ex. |
| bl._a. | bl.a. |
| bl_a | bl_a |
| d._v._s. | d.v.s. |
| d_v_s | d_v_s |
| e._d. | e.d. |
| e_d | e_d |
| f._n. | f.n. |
| f_n | f_n |
| fr._o._m. | fr.o.m. |
| m._fl. | m.fl. |
| m_fl | m_fl |
| m._m. | m.m. |
| m_m | m_m |
| o._s._v. | o.s.v. |
| o_s_v | o_s_v |
| s._k. | s.k. |
| s_k | s_k |
| t._ex | t.ex |
| t._ex. | t.ex. |
| t_ex | t_ex |
| t._h. | t.h. |
| t._o._m. | t.o.m. |
| t_o_m | t_o_m |
| t._v. | t.v. |
| t_v | t_v |

Due to the concatenation of tokens found in the abbreviation list above, the original word numbering for these cases will contain gaps. Therefore, a new sequential token numbering is added for the whole data set. Further, the abbreviation recieves the lex pos feature of the first token in the expression, i.e. the ID tags for the other tokens belonging to the abbreviation are removed.

Example of mamba items before processing:

```
<row wordnr="017" st="07" gm="014" ms="015" dk="" us="" syntax="TAPR" textnr="P101" lex="PR">fr.</row>
<row wordnr="018" st="07" gm="014" ms="015" dk="" us="" syntax="TAPR" textnr="P101" lex="ID">o.</row>
<row wordnr="019" st="07" gm="014" ms="015" dk="" us="" syntax="TAPR" textnr="P101" lex="ID">m.</row>
<row wordnr="020" st="07" gm="014" ms="015" dk="" us="" syntax="TADT" textnr="P101" lex="PODP">den</row>
```

Example of mamba items after processing:

```
<row n="8" wordnr="017" st="07" gm="014" ms="015" dk="" us="" syntax="TAPR" textnr="P101" lex="PR">fr._o._m.</row>
<row n="9" wordnr="020" st="07" gm="014" ms="015" dk="" us="" syntax="TADT" textnr="P101" lex="PODP">den</row>
```

Noteworthy is that we found two errors in the annotation of abbreviations in Talbanken: bl. a. på", number P10712052222, and bl a på", number P2049154952.

# Referenser


[Ejerhed *et* al. 1992] Ejerhed, E., Källgren, G., Wennstedt, O. and Åström, M. (1992) The Linguistic Annotation System of the Stockholm-Umeå Project Department of General Linguistics, University of Umeå.

[Källgren *et* al. 2006] Källgren, G. Gustafson-Capkova, S. and Hartmann, B. (2006) Stockholm Umeå Corpus 2.0 (SUC2.0) Department of Linguistics, Stockholm University, Stockholm, Sweden.

[SUC1.0 1997] SUC 1.0 Stockholm Umeå Corpus, Version 1.0, (1997) SUC. Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University ISBN: 91–7191–348–3

[SUC2.0 2006] SUC 2.0 Stockholm Umeå Corpus, Version 2.0, (2006) SUC. Department of Linguistics, Stockholm University ISBN: ?

[Teleman 1974] Teleman, U. (1974) Manual för grammatisk beskrivning av talad och skriven svenska (Mamba) Lund University

[Westman 1974] Teleman, U. (1974) Bruksprosa Lund University


**Appendix: Multiword expressions in Talbanken**

The abbreviations are annotated as *ABBR and the wrongly extracted abbreviations (due to the erroneus annotation as ID pos tag in the original file) are marked as *WRONG. Also, some expressions have a comment added to it marked as #.

1000_husmödrar_om_hemarbetet
10_f
1._1.
1_/_1
11_/_20
12_/_20
13_/_20
1967_/_68
1974_/_75
222_Stockholmspojkar
222_stockholmspojkar
295_ff
30_ff
3_-_4
6_X_6
AB_Air_Union
A._I._Rabin
ATP_de_första_åren
Ab_Elof_Malmberg
Agnes_Varda
Aldus_/_Bonniers
Allmänna_Förlaget
Allt_efter_som
Allt_fortfarande
American_Journal_of_Orthopsychiatry
Angeläget_nu
Anna-Britta_Hellgren
Anna-Lisa_Kälvestam
Anna-Lisa_Kälvestens
Annika_Baude
Att_bygga_ekonomisk_trygghet_på_känslor
Barbro_Backberger
Bengt_Silfverberg
Berl_Kutschinsky
Bertil_Schmidt
Birger_Rohlin
Birger_Svensson
Björn_Gillberg
Björn_Gillbergs
Björn_Runeborgs
*ABBR Bl._a.
*ABBR Bl_a

Bland_annat
Boiling_Water_Reactor
Bortsett_från
Brita_Wigforss
Bromma_gymnasium
C_E_Johansson
Carin_Boalt
Carin_Boalts
Central_Park
Centrala_folkbokförings-_och_uppbördsnämnden
Claes_Folkeson
Claes_Varenius
Curt_Steffan_Giesecke
Dagens_Nyheter
Daniel_Horn
De_Sex
De_här
Den_här
Den_permanente
Den_som_bestämmer_mest_hemma_hos_oss
De_sex
De_särskilda_brotten
De_ständiga_representanternas_kommitte
Det_här
Detta_till_trots
Dina_förmåner
Därför_att
Dx_-
E._W._Burgess
Efter_hand
En_del
En_i_taget
En_och_annan
Ett_par
Ett_tag
European_Economic_Community
European_Free_Trade_Association
Europeiska_arbetsmarknadsverket
Europeiska_centralbanken
Europeiska_forskningsnämnden
Europeiska_förenade_fackföreningsrörelsen
Europeiska_institutet_för_företagsutbildning
Europeiska_investeringsbanken

Europeiska_lantbruksnämnden
Europeiska_patentverket
Familjen_i_samhället
Fängelset_som_vårdform
Food_and_Agriculture_Organization_of_the_United_Nations
Framför_allt
Frances_Westin
För_att
För_det_andra
För_det_första
Förenta_staterna
*ABBR Fr_o_m
Förr_eller_senare
Försäkringsbolaget_Trygg
Först_och_främst
Förteckning_över_postanstalter_i_Sverige
För_övrigt
G_H_T
Gamla_testamentet
George_Murdock
Georges_Papy
Gertrud_Schyl-bjurman
Gordon_Rattray_Taylor
Göran_Bergman
Göran_Löfroth
Göran_Ryding
Gösta_Lindebo
Gudrun_Werkström
Gunnar_Boalt
Gunnar_Hedlund
Gunnar_Helen
Gunnar_Hägglöf
Gunnar_Hultgren
Gunnar_Sträng
Gustav_Jonsson
Gustav_Jonssons
Ha_ha
Hans-jörgen_Hansen
Hemmafru_i_skatteskruv
Hisings_Backa
Håkan_Ohlssons
Ho_ho
Hotell-_och_Restauranganställdas_förbund

Hur_pass
I_anslutning_till
I_bästa_fall
Icke_desto_mindre
I_dag
I_de_fall
I_fråga_om
I_går
I_jämförelse_med
I_morgon
Indiska_oceanen
Infants_and_Children_under_Conditions_of_'_Intermittent_'_Mothering_in_the_Kibbutz
Ingemund_Bengtsson
Ingmar_H:son_Klockhoff
Ingrid_Gunneson
Ingrid_Sjöstrand
I_och_med
I_år
I_regel
I_runda_tal
I_runt_tal
I_så_fall
I_själva_verket
I_skydd_av
I_stället
I_stället_för
I_varje_fall
I_vintras
I_övrigt
J._Bowlby
Jack_Adams-Ray
Jan_Trost
January_1970
J_de_fall
Jean_Piaget
Jerome_Bayless
Jesus_Kristus
John_Erik_Boork
John_Lind
Jo_visst
Just_det
K._Elmhorn
Kajsa_Ohrlander

Kajsa_Ohrlanders
Katolska_kyrkan
Kibbutzchildren_-_Researchfindings_to_Date
Kjell-Olof_Feldt
Kommunalt_bostadstillägg
Konsumentinstitutet_meddelar
Kriminalpolitiska_åtgärder
Kriminalvård_på_anstalt
Krister_Wickman
Kungl._Maj:t
Kungl_Maj:t
Kvarnvikens_Båtsällskap
Kvarnvikens_båtsällskap
Kvinnor_som_slavar
Lars-erik_Håkansson
Lokala_skattemyndigheten
Ludvig_Jönsson
Magnus_Gernes
Margaret_Meads
Maria_Montessoris
Martin_Luther
Med_hjälp_av
Medicinsk_bildtolk
Med_tanke_på
Mer_än
Mindre_än
Min_lycka
Mt._19_:_10_ff.
National_Clearinghouse_for_Smoking_and_Health
New_York
New_thinking_in_school_mathematics
Något_av
Något_som
Niko_Tinbergen
Nils_Ericsson
Nils-gustav_Gejvall
Nora_Stationsgatan
Nordiska_kommitten_för_modernisering_av_matematikundervisningen
Norra_Botkyrka
North_Atlantic_Treaty_Organization
När_det_gäller
Nya_Testamentet
Nya_Zeeland

Nya_testamentet
Oljeprospektering_AB
Olof_Magne
Olof_Palme
Oscar_Öqvist
Oscars_församling
Oskar_Öqvist
Ostkustfisk_Centralförening
Otto_Wagner
Ove_Hemer
Paul-Henri_Spaak
Per-Ola_Larsson
Per_capita
Per-erik_Almgren
På_grund_av
På_grundval_av
Poste_restante
Press_och_Information
Pressurized_Water_Reactor
På_så_sätt
Ragnar_Muller
Ragnar_Sohlman
Röda_Korset
Red_Top
Rose_Kennedy
Rose_Kennedys
S_-
Samhället_och_ungdomsbrottslingarna
Samtidigt_som
Så_att
Södra_Kyrkogatan
Sergels_torg
Så_här
Självdeklarerad_brottslighet_bland_skolbarn
Så_länge
Sociala_huset
Socialpolitiken_i_ett_internationellt_perspektiv
Som_bekant
Så_småningom
Så_snart
S:t_Eriksgatan
S:t_Lars_sjukhus
Standard_Oil_of_California

Statens_Naturvårdsverk
Statens_institut_för_Folkhälsan
Statens_pris-_och_kartellnämnd
Statens_veterinärmedicinska_anstalt
Statistisk_Årsbok
Statistisk_årsbok
Sten_Cronqvist
Sten_Sjöholm
Stig_Larsson
Stilla_oceanen
Ständiga_representanternas_kommitte
Stora_Katekesen
Sven_Backlund
Sven_Weden
Svenska_Dagbladets
Svenska_FN-förbundets
Svenska_ostkustfiskarnas_centralförbund
Sveriges_Meterologiska_och_Hydrologiska_Institut
Sveriges_Radios_förlag
Sverker_Åström
Swedish_International_Development_Authority
Tack_vare
*ABBR T._ex.
Thord_Erasmie
Thure_Andersson
Till_exempel
Till_och_med
Till_skillnad_från
Till_skillnad_mot
Till_storms_mot_äktenskapet
Tjäna_på_att_veta_om_leksaker
Tom_Hardt
Tomas_Hedqvist
Torsten_Nilsson
Trots_att
Ulla_Hasselquist
Undan_för_undan
Under_motorhuven
Undervisning_eller_undergång
Unesco_Courier
Vad_nu_Gud_har_sammanfogat_...
Var_fjärde
Var_och_en

Vdn_Fakta
Veckans_affärer
Vem_som_helst
Vid_sidan_av
Västra_Frölunda
Walter_Scheel
Z_P_Dienes
a._a. #namninitialer P405004135963 i kontexten "Jönsson A.A., sid."
afUgglas_ägo
aldrig_så
allt_efter
allt_eftersom
alltefter_som
allt_fler
allt_flera
allt_fortfarande
allt_mer
allt_mera
Å_andra_sidan
å_andra_sidan
annat_än
a_priori
av_allt_att_döma
av_inre_ursprung
barns_skull
barnuppfostran_för_slags
både_till
beroende_på
biologiska_klockor
*ABBR bl._a.
*ABBR bl_a
bland_andra
bland_annat
*WRONG! bl._a._på
*WRONG! bl_a_på
bortsett_från
brutaliteter_som_helst
companionship_family
consensus_facit_nuptiam
corn_flakes
council_house
dag_och_natt
då_det_gäller

decennier_sedan
de_här
å_den_andra
den_här
det_där
det_för
det_här
det_samma
det_vill_säga
då_och_då
därför_att
drilling_manager
*ABBR d._v._s.
*ABBR d_v_s
*ABBR e._d.
*ABBR e_d
efter_det_att
efter_hand
efter_hand_som
å_ena_sidan
en_del
en_eller_annan
en_halv
en_i_taget
en_och_en
en_och_en_halv
en_och_samma
ett_dugg
ett_eller_annat
ett_halvt
ett_och_ett_halvt
ett_par
fall_som_helst
fil_lic
fler_än
*ABBR f._n.
*ABBR f_n
fortplantningens_skull
för_all_del
för_alltid
framför_allt
fram_och_tillbaka
för_att

för_det_mesta
före_detta
från_det
från_och_med
från_och_med_det
för_närvarande
*ABBR fr._o._m.
fr_o_m
förr_eller_senare
förr_i_världen
för_sed
först_och_främst
förutsatt_att
för_övrigt
fullt_ut
gång_på_gång
halv_tre
hand_i_hand
helt_enkelt
helt_och_hållet
högre_än
hålkorten_-_valsedlarna
hur_som_helst
i_alla_fall
i_alla_händelser
i_allmänhet
i_all_världen
i_andra_hand
i_anslutning_till
i_dag
i_dagsläget
i_den_mån_,_som
i_enlighet_med
i_fatt
i_fjol
i_följd
i_form_av
i_fortsättningen
i_förbindelse_med
i_förening_med
i_fråga
i_fråga_om
ifråga_om

i_förgrunden
i_förhållande_till
i_första_hand
i_förtid
i_förväg
i_funktion
i_gång
i_går
i_grunden
i_hög
i_hög_grad
i_höst
i_huvudsak
i_jämförelse_med
i_kläm
i_konflikt
i_kontakt
i_kraft
i_längden
i_mån_av
i_morse
i_motsats_till
i_natt
inget_annat_än
i_någon_mån
innehållet_beträffar
inom_kort
inom_ramen_för
i_norr
i_närheten_av
inte_minst
i_och_för_sej
i_och_för_sig
i_och_med
i_onödan
i_ordning
i_princip
i_proportion
i_proportion_till
i_år
i_regel
i_riktning_mot
i_runt_tal

i_sak
i_samarbete_med
i_samband_med
i_söder
i_så_fall
i_sin_tur
i_själva_verket
i_skuggan
i_sämsta_fall
i_så_måtto
i_söndags
i_säng
i_somras
i_stället
i_stället_för
istället_för
i_stånd
i_stort
i_stort_sett
i_synnerhet
i_taget
i_takt_med
i_tid
i_undantagsfall
i_underkant_av
i_vad_mån
i_valet_och_kvalet
i_varje_fall
i_viss_mån
i_vår
i_våras
i_övrigt
jämfört_med
Åke_Nilsson
kol-_och_stålunionen
Äktenskapet_fri_samlevnad
lägre_än
lika_med
länge_sedan
långt_ifrån
låt_vara
lustigt_nog
med_andra_ord

med_avseende_på
med_fog
med_hjälp_av
med_hänsyn
med_hänsyn_till
med_hänvisning_till
med_ledning_av
med_lic
med_mera
med_sikte_på
med_tanke_på
med_undantag_för
med_utgångspunkt_av
med_utgångspunkt_i
mer_eller_mindre
mer_än
mer_och_mer
*ABBR m._fl.
*ABBR m_fl
mindre_än
mindre_och_mindre
minst_av_allt
*ABBR m._m.
*ABBR m_m
mycket_väl
något_av
något_som
några_som_helst
ännu_så_länge
nog_så
nära_nog
när_det_gällde
när_det_gäller
när_det_gällt
när_som_helst
än_så_länge
oberoende_av
och_på
och_så_vidare
orsaks_skull
*ABBR o._s._v.
*ABBR o_s_v
ovan_nämnda

på_allvar
på_basis_av
på_det
på_det_klara
på_efterkälken
på_egen_hand
på_ett_ungefär
på_för
på_frågan
på_förhand
på_gång
på_grund
på_grund_av
på_grundval_av
på_längre_sikt
på_nytt
praktiskt_taget
pressure_groups
på_rygg
på_sikt
på_sina_håll
på_sin_höjd
på_sistone
på_spel
på_så_sätt
på_tal
på_tal_om
på_tok
på_väg
på_vippen
relativt_sett
rent_av
är_för_något
år_från_år
år_sedan
år_sen
rättare_sagt
rätt_så
(_s_)
samtidigt_med_att
samtidigt_som
så_att
så_att_säga

så_borgerligt-individualistiska
sådana_här
sådan_här
så_där
sent_omsider
sex_gånger_sex
så_fan_heller
så_fort
så_gott_som
så_här
så_här_års
å_sin_sida
sist_berörda
sist_men_inte_minst
situationer_som_helst
*ABBR s._k.
*ABBR s_k
så_kallad
säkerhets_skull
så_länge
så_långt
sån_här
som_bekant
som_regel
som_synes
som_vanligt
så_pass
så_småningom
steg_för_steg
stick_i_stäv
strängt_taget
såväl_som
så_värst
syndens_skull
synen_beträffar
tack_vare
*ABBR t._ex
*ABBR t._ex.
*ABBR t_ex
*ABBR t._h.
tid_sedan
till_användning
till_buds

till_exempel
till_följd_av
till_förmån_för
till_fullo
till_godo
till_hands
till_havs
till_höger
till_höger_och_vänster
till_höger_om
till_låns
till_och_med
till_ro
till_rätta
till_sist
till_skillnad_mot
till_slut
till_sängs
till_stånd
tills_vidare
till_synes
till_tjänst
till_uttryck
till_vänster
till_vänster_om
*ABBR t._o._m.
*ABBR t_o_m
totalt_sett
trots_allt
trots_att
*ABBR t._v.
*ABBR t_v
under_alla_förhållanden
under_debatt
under_det_att
under_förutsättning_att
under_loppet_av
upp_och_nedvänt
upp_till
ur_famn_i_famn
ur_spel
utan_avseende_på
utan_hänsyn_till

utan_tanke_på
utan_tvekan
utan_vidare
vad_beträffar
vad_det_gäller
vad_gäller
vad_mera_är
vad_som_helst
vara_för
vare_sig
var_femte
var_fjärde
var_för_sig
var_för_slags
var_och_en
var_sin
var_sitt
var_som_helst
var_tionde
vart_och_ett
var_tredje
vart_som_helst
vem_som_helst
Även_om
även_om
överallt_.
över_huvud
över_huvud_taget
överhuvud_taget
Övervakningsnämndernas_organisation_och_verksamhet
vice_versa
vid_det_laget
vid_handen
vid_sidan_av
vid_sidan_om
vill_säja
yrkesarbete_som_helst