Tsez –

# EDITS MADE to INDIVIDUAL LINES in DATA

The lines below contain a **single word accidentally separated** by extra spaces or else **multiple words (or glosses or lexical category tags) were united**, usually because they were missing a tab ('\t) in the segmentation or gloss line. I confirmed this by finding the same word elsewhere in the data. I **edited** them in ddo-all-txts-examples01.csv

43858, dido1241:  
"Boƛƛ’o zaman bik’inosi, kidbä neła nełdaɣor eniw yižin.",  
boƛƛ’o\tzaman\tb-ik’i-nosi\tkid-a\tneła-a\tneła-daɣor\t**eniw y-iži-n**,  
middle\ttime\tIII-go-ANT.CVB\tgirl-ERG\tDEM1.IISG.OBL-ERG\tDEM1.IISG.OBL-APUD.VERS\**tmother\tII-lead-PST.UNW**,  
adv n3 pref-v-vsuf n2-nsuf pron-nsuf pron-nsuf n2 pref-v-vsuf"

14-170, dido1241:   
[**'hay i wele'**, 'howži', 'oxi', 'mi', 't’ay', 'di-s', 'xediw', 'ay-näy', 'exur', 'nesi-a', 'mi']  
**so.thenanlook.out.IMP**\tnow\trun.away\tyou\tfrom.here\tme-GEN1\thusband\tcome-CND.CVB\tkill\tDEM1.ISG.OBL-ERG\tyou >>  
**hay\ti\twele**\thowži\toxi\tmi\tt’ay\tdi-s\txediw\tay-näy\texur\tnesi-a\tmi,  
**so.then\tand\tlook.out.IMP**\tnow\trun.away\tyou\tfrom.here\tme-GEN1\thusband\tcome-CND.CVB\tkill\tDEM1.ISG.OBL-ERG\tyou  
**excl exexcl** adv v pron adv pron-nsuf n1 v-vsuf v pron-nsuf pron

62-3, dido1241:   
Hemeła ɣudƛ’ot’a q’ˤuya ulkał ičixosi Muħamadƛin ci yołäsi **muʔmin žek’u elor** aħnayx zown.   
['hemeła', 'ɣudi-ƛ’-t’a', 'q’ˤuya', 'ulka-ł', 'iči-xosi', 'Muħamad-ƛin', 'ci', 'yoł-asi', **'muʔminžek’uelo-r'**, 'aħnad-x', 'zow-n']   
DEM4.IISG.OBL\tday-SUPER.ESS-DISTR\tother\tcountry-CONT.ESS\tlive-PRS.PRT\tMuhammad-QUOT\tname\tbe-RES.PRT\t**piousmanthere-**LAT\twalk.around-IPFV.CVB\tbe.NPRS-PST.UNW  
"pron n4-nsuf-suf n n3-nsuf v-vsuf n1-suf n3 v-vsuf **adjn1adv-nsuf** v-vsuf v-vsuf"

62-49, dido1241:   
"Sidä eƛin, yisä däqay moƛał ʕazar ɣuruš yoɣäsi zows, hoboy ža yisä **xizor neƛxanuƛin**."   
['sida-a', 'eƛi-n', 'yisi-a', 'di-qay', 'moƛu-ł', 'ʕazar', 'ɣˤuruš', 'y-oɣ-asi', 'zow-s', 'hoboy', 'ža', 'yisi-a', **'xizorneƛ-x-anu-ƛin'**]   
num-nsuf v-vsuf pron-nsuf pron-nsuf n2-nsuf num n2 pref-v-vsuf v-vsuf adv pron pron-nsuf **advv-vsuf-vsuf-suf"**

76-28, dido1241,  
"Xizyo zeyä nesiqor eƛin, sidaɣ qoqä teł łaces gulubi yoł, mi ik’in elo q’ˤim **ritäruntow, ža qoq baɣˤiłno näx**.",   
['xizyo', 'zey-a', 'nesi-qor', 'eƛi-n', 'sidaɣ', 'qoq-a', 'teł', 'łaci-s', 'gulu-bi', 'yoł', 'mi', 'ik’i-n', 'elo', 'q’ˤim', **'r-iti-run-towžaqoqb-ˤaɣˤi-ł-nnex**']  
afterwards\tbear-ERG\tDEM1.ISG.OBL-POSS.LAT\tsay-PFV.CVB\tin.one.place\tcliff-IN.ESS\tinside\twind-GEN1\thorse-PL\tbe**\tyou\tgo-PFV.CVB\tthere\thead\tIV.PL-touch-IMM.ANT.CVB-EMPH** **HDEM1.SGcliffIII-open-POT-PFV.CVBcome**adv n3-nsuf pron-nsuf v-vsuf adv n4-nsuf post n4-nsuf n3-nsuf v pron v-vsuf adv n4 **pref-v-vsuf-sufpronn4pref-v-vsuf-vsufv**

78-8, dido1241,  
"Hudayzikon uži bełiƛ’ ik’in, qˤayn **č’aq’ˤirno**   **xizor utin** ayn.",  
hudayziko-n\tuži\tbełi-ƛ’\tik’i-n\tqˤay-n\tč’aq’ˤi-r**-n**\**txizor\tuti**-n\tay-n,  
the.next.day-TOP\tson\thunting-SUPER.ESS\tgo-PFV.CVB\thare-TOP\thit-CAUS-**PFV.CVB**\**tback\tturn.back**-PFV.CVB\tcome-PFV.CVB,  
"adv-suf n1 n3-nsuf v-vsuf n3-suf v-vsuf-**vsuf adv v**-vsuf v-vsuf"

62-49,dido1241,  
"Sidä eƛin, yisä däqay moƛał ʕazar ɣuruš yoɣäsi zows, hoboy ža yisä **xizor ne**ƛxanuƛin.",  
sida-a\teƛi-n\tyisi-a\tdi-qay\tmoƛu-ł\tʕazar\tɣˤuruš\ty-oɣ-asi\tzow-s\thoboy\tža\tyisi-a\t**xizor\tneƛ**-x-anu-ƛin,  
one.OBL-IN.ESS\tsay-PST.UNW\tDEM2.ISG.OBL-ERG\tme-POSS.ABL\tdream-CONT.ESS\tthousand\truble\tII-take.away-RES.PRT\tbe.NPRS-PST.WIT\tnow\tDEM1.SG\tDEM2.ISG.OBL-ERG\**tback\tgive**-IPFV.CVB-NEG-QUOT,  
 "num-nsuf v-vsuf pron-nsuf pron-nsuf n2-nsuf num n2 pref-v-vsuf v-vsuf adv pron pron-nsuf **adv v**-vsuf-vsuf-suf"

ln 29: 43858,dido1241,  
"Boƛƛ’o zaman bik’inosi, kidbä neła nełdaɣor eniw yižin.",  
boƛƛ’o\tzaman\tb-ik’i-nosi\tkid-a\tneła-a\tneła-daɣor\teniw\**ty-iži-n,**  
middle\ttime\tIII-go-ANT.CVB\tgirl-ERG\tDEM1.IISG.OBL-ERG\tDEM1.IISG.OBL-APUD.VERS\tmotherII\t**lead-PST.UNW**  
adv n3 pref-v-vsuf n2-nsuf pron-nsuf pron-nsuf **n2 pref**-v-vsuf"

This line was **DELETED** because there was no segmentation, glosses, etc.

66-238,dido1241,  
Qu nukar egirno xanä nesir hawšinaw rigłi radiru žedu xexbi goƛ’anix.,qu nukar egirno xanä nesir hawšinaw rigłi radiru žedu xexbi goƛ’anix,qu nukar egirno xanä nesir hawšinaw rigłi radiru žedu xexbi goƛ’anix,,stan1293,,66,,qu nukar egirno xanä nesir hawšinaw rigłi radiru žedu xexbi goƛ’anix

These lines had a **MISSING WORD.** Iinserted it in the file (based on the other lines and other instances of the word in the data.)

75-119, dido1241   
['iysƛin', **'eƛin'**, 'kidbez', **'**babiy**a'**, 'ciqo', 'imxa', 'yäč’iruni', 'gonda', 'teł', 'zeyes', 'lidaƛ', 'waħo', 'uqˤłäsi', 'mi', 'elur', 'ukaysƛin.']   
[**'iy-s-ƛin', 'kid-z'**, 'babiw-a', 'ciq', 'imxu-a', 'y-ič’i-ru-ni', 'gondu-a', 'teł', 'zey-s', 'lidu-ƛ', 'waħo', 'uqˤeł-asi', 'mi', 'elu-r', 'ukad-s-ƛin']  
**know-PST.WIT-QUOT\tdaughter-GEN2**\tfather-ER**G**\tforest\tcave-IN.ESS\tII-dig-PST.PRT-DEF\tbig.hole-IN.ESS\tinside\tbear-GEN1\tpaw-SUB.ESS\tbelow\thide-RES.PRT\tyou\twe(I).OBL-LAT\tsee-PST.WIT-QUOT  
**v-vsuf-suf n2-nsuf** n1-nsuf n4 n2-nsuf pref-v-vsuf-suf n2-nsuf post n3-nsuf n2-nsuf post v-vsuf pron pron-nsuf v-vsuf-suf  
""We know, **said** the daughter's father, we saw you in the forest in the cave in the big hole dug by bear under his paw.""

78-8, dido1241   
Hudayzikon uži bełiƛ’ ik’in, qˤayn č’aq’ˤirno **xizor** utin ayn.  
['hudayzikon', 'uži', 'bełiƛ’', 'ik’in', 'qˤayn', 'č’aq’ˤirno', **'xizor'**, 'utin', 'ayn.']   
['hudayziko-n', 'uži', 'bełi-ƛ’', 'ik’i-n', 'qˤay-n', **'č’aq’ˤi-r-n oruti-n'**, **'ay-n'**]  
the.next.day-TOP\tson\thunting-SUPER.ESS\tgo-PFV.CVB\thare-TOP\**thit-CAUS-**PFV.CVB kturn.back-PFV.CVB\tcome-PFV.CVB  
adv-suf n1 n3-nsuf v-vsuf n3-suf **v-vsuf-vsufv-vsuf** v-vsuf

\*\*\*

**An extra morpheme** in the segmentation line**.** This occurred in the example below and 19 other instances of this lemma (**'banquet; fun'**) The morpheme was combined the file.

30-11, dido1241,  
"Hemedur uynon **rok’uɣʷeymoƛ’** hunix rik’iƛ’oräy, žedar bexuräsi boc’i besun.",  
hemedur\tuyno-n\**trok’u-ɣʷey-ƛ’**\thuni-x\tr-ik’i-ƛ’oräy\tžeda-ar\tb-exur-asi\tboc’i\tb-esu-n,  
so\tfour-TOP\**tfun-SUPER.ESS**\tway-AD.ESS\tIII.PL-go-SIM.CVB\tDEM1.IIPL.OBL-IN.LAT\tIII-kill-RES.PRT\twolf\tIII-find-PST.UNW,

# CLEANING UP LEXICAL CATEGORY TAGS

This is the original set of unique morpheme lexical category tags. The bolded and crossed out ones were edited in the file. In most cases, they needed to be separated by a tab instead of space(s).

{'adj', **~~'adj n1 adv'~~,** 'adv', **'~~adv v'~~**, 'comp', 'conj', 'deriv', 'excl', **~~'excl '~~, ~~'excl excl excl'~~,**  'interj', 'n', 'n1', **~~'n1.pl~~', 'n1pl'**, 'n2', **~~'n2 pref'~~,** 'n2pl', 'n3', **~~'n3 '~~,** 'n3pl', 'n4', 'nsuf', 'num', 'numsuf', 'post', 'pref', 'pron', **~~'pron '~~, ~~'pron '~~**, **'su',** 'suf', **~~' v '~~**, 'v', 'vsuf', **~~'vsuf adv v'~~, ~~'vsuf\\'~~**}

The ones below, however, had to be changed.

**'n1.pl'** 🡪 **'n1pl'  
'su' 🡪 'suf'**

# Changes during preprocessing

There are 689 hyphenated (compound?) words (672 not counting words ending with hyphens, see below) in the original text that get treated as separate morphemes on the segmented line, e.g.

**'ɣˤana-xediw'   
'ɣˤana', 'xediw'  
'married.couple'**

These are replaced with ~ in during processing but not in the file. It is not a simple replacement because many also have a morpheme division and 133 of them have different letters before and/or after the compounding hyphen in the underlying morpheme line. Most of those 133 will not be included in the model, e.g.

'k’e**k’-b**ik’erxot’a'  
'k’e**k’-i**k’-r-x-t’a'

The below I edited in the file:

'make.red-hot' 🡪 make.red.hot – 5 instances.

son-in-law in glosses 🡪 son~in~law – 1 instance.

**she-goat** in glosses 🡪 **she.goat** - 24 instances.

he-goat 🡪 he.goat - 12 instances.

Not sure why these words end in hyphens? (my guesses are right of arrows) In preprocessing the hyphens will be removed and the words treated as separate, single words.

ln223 ['reže-', 'r-eže-', ['IV', 'big'], ['pref', 'adj']] 🡨 self-edit, delete

ln413 ['yił-', 'yiła', ['DEM2.IISG.OBL'], ['pron']] 🡨 ditto??

ln805 ['eni-', 'eni', ['mother'], ['n2']] 🡨 compound with next word, should combine

ln831 ['eni-', 'eni', ['mother'], ['n2']] 🡨 ditto

ln1153 ['eni-', 'eni', ['mother'], ['n2']] 🡨 ditto

ln1160 ['eƛi-', 'eƛi', ['say'], ['v']] 🡨 self-edit, delete

ln2731 ['eni-', 'eni', ['mother'], ['n2']] 🡨 compound with next word, combine

ln2791 ['his-', 'his', ['spring.wool'], ['n4']] 🡨 compound??

ln3017 ['ħaƛu-', 'ħaƛu', ['drink'], ['v']] 🡨 compound/serial??

ln4787 ['bezu-', 'b-ezu-', ['III', 'look'], ['pref', 'v']] 🡨 ???

ln4795 ['žek’-', 'žek’', ['beat'], ['v']] 🡨 compound/serial?

ln4797 ['ħilla-', 'ħilla', ['trickery'], ['n3']] 🡨 compound?

ln4807 ['babiw-', 'babiw', ['father'], ['n1']] 🡨 compound?

ln4824 ['bezu-', 'b-ezu-', ['III', 'look'], ['pref', 'v']] 🡨 ??? cf ln 4787

ln4882 ['caƛi-', 'caƛi', ['throw'], ['v']] 🡨 compound/serial?

ln4891 ['elo-', 'elo', ['there'], ['adv']] 🡨 compound, combine