Recovering Text from Endangered Languages Corrupted PDF documents

Nicolas Stefanovitch

European Commission, Joint Research Center, Ispra, Italy

ComputEL-5 Workshop
May 26-27 2022

Universal Declaration of Human Rights in Nivkh (niv)

Қатьгун сик правоғун Декларация
Генеральная Ассамблея резолюция 217A (III)
10 к`лолоң 1948 ань ӽавыл ғера
провозглашайра ӽад

Нивң қ`атьгун ӽара, чуғун ӽара сик намадивңчоғр̆ иввут т`оҳтта, адяй сик п`ңафқ-ңафқ ңазин правоғун ивңы-худ куғытнд ӽара салғанӽунвд ӽара сикғун мир ӽара; адяй пиған сикак к`икр̆ п`ид-ӽаымди қ`оғл ивкғун нивгун правогун лақ қаврта ӽа-ӽаңы варвар актғундоҳ вылңута, худғундоҳ нивгун сик макыр қ`оғл легу осқита ӽагут урла қ`оғл ивң нивгунаҳ п`ңалзайгута; адяй айф тымди мир айғай – ху мирух нивгун кулғытң туғс ӽара аҳтнд ӽара ивңы адяй к`лунд ӽара ӽекинд ӽара ағзута – ӽудғун нивгун язра ныйнынд; адяй нивгун ӽекиr̆киr̆ п`синрукғундоҳ озн вопун ухмуна ӽа ғавргуйныr̆киr̆ нивгун право властьғун торкиr̆ ытңуна; адяй қ`атьгун п`ңафқ-ңафқроҳ ңафқңалагуйнын ихмына ӽагуйн; адяй объединенных наций қ`атьгун п`Уставух адяй қ`атьгун правоғун аҳтхымлыта, нивгун п`и намад ивд яймта азмытьғундиң р̆аңғундиң салғат ӽумта ӽатот социальный прогресс ӽагуйныңа адяй нивгун малғо куғытнд ивт ӽунвд ургута; адяй государствоғун – членгун Организация объединенных наций ройныт т`оҳтта,

:àòüãóí ñèê ïðàâîâ4óí Äåêëàðàöèÿ
Ãåíåðàëüíàÿ Àññàìáëåÿ ðåçîëþöèÿ 217A (III)
10 ê`ëîëî2 1948 àíü 1àâûë 4åðà
ïðîâîçãëàøàéðà 1àä

Íèâ2 6`àòüãóí 1àðà, ÷ó4óí 1àðà ñèê íàìàäèâ2÷î43 èââóò ò`î5òòà, àäÿé ñèê ï`2àô6-2àô6 2àçèí ïðàâî4óí èâ2û-õóä êó4ûòíä 1àðà ñàë8àí1óíâä 1àðà ñèê4óí ìèð 1àðà; àäÿé ïè8àí ñèêàê ê`èê3 ï`èä-1àûìäè 6`î8ë èâê4óí íèâãóí ïðàâîâóí ëà6 6àâðòà 1à-1à2û âàðâàð àêò4óíäî5 âûë2óòà, õóä4óíäî5 íèâãóí ñèê ìàêûð 6`î8ë ëåãó îñ6èòà 1àãóò óðëà 6`î8ë èâ2 íèâãóíà5 ï`2àëçàéãóòà; àäÿé àéô òûìäè ìèð àé8àé – õó ìèðóõ íèâãóí êóë4ûò2 òó4ñ 1àðà à5òíä 1àðà èâ2û àäÿé ê`ëóíä 1àðà 1åêèíä 1àðà à4çóòà – 1óä4óí íèâãóí ÿçðà íûéíûíä; àäÿé íèâãóí 1åêè3êè3 ï`ñè2ðóê4óíäî5 îçí âîïóí óõìóíà 1à 9àâðãóéíû3êè3 íèâãóí ïðàî âëàñòü4óí òîðêè3 ûò2óíà; àäÿé 6`àòüãóí ï`2àô6-2àô6ðî5 2àô62àëàãóéíûí èõìûíà 1àãóéí; àäÿé îáúåäèíåííûõ íàöèé 6`àòüãóí ï`Óñòàâóõ àäÿé 6`àòüãóí ïðàâî4óí à5òõûìëûòà, íèâãóí ï`è íàìàä èâä ÿéìòà àçìûòü4óíäè2 3à294óíäè2 ñàë8àò 1óìòà 1àòîò ñîöèàëüíûé ïðîãðåññ 1àãóéíû2à àäÿé íèâãóí ìàë8î êó4ûòíä èâò 1óíâä óðãóòà; àäÿé ãîñóäàðñòâî4óí – ÷ëåíãóí Îðãàíèçàöèÿ îáúåäèíåííûõ íàöèé ðîéíûò ò`î5òòà,

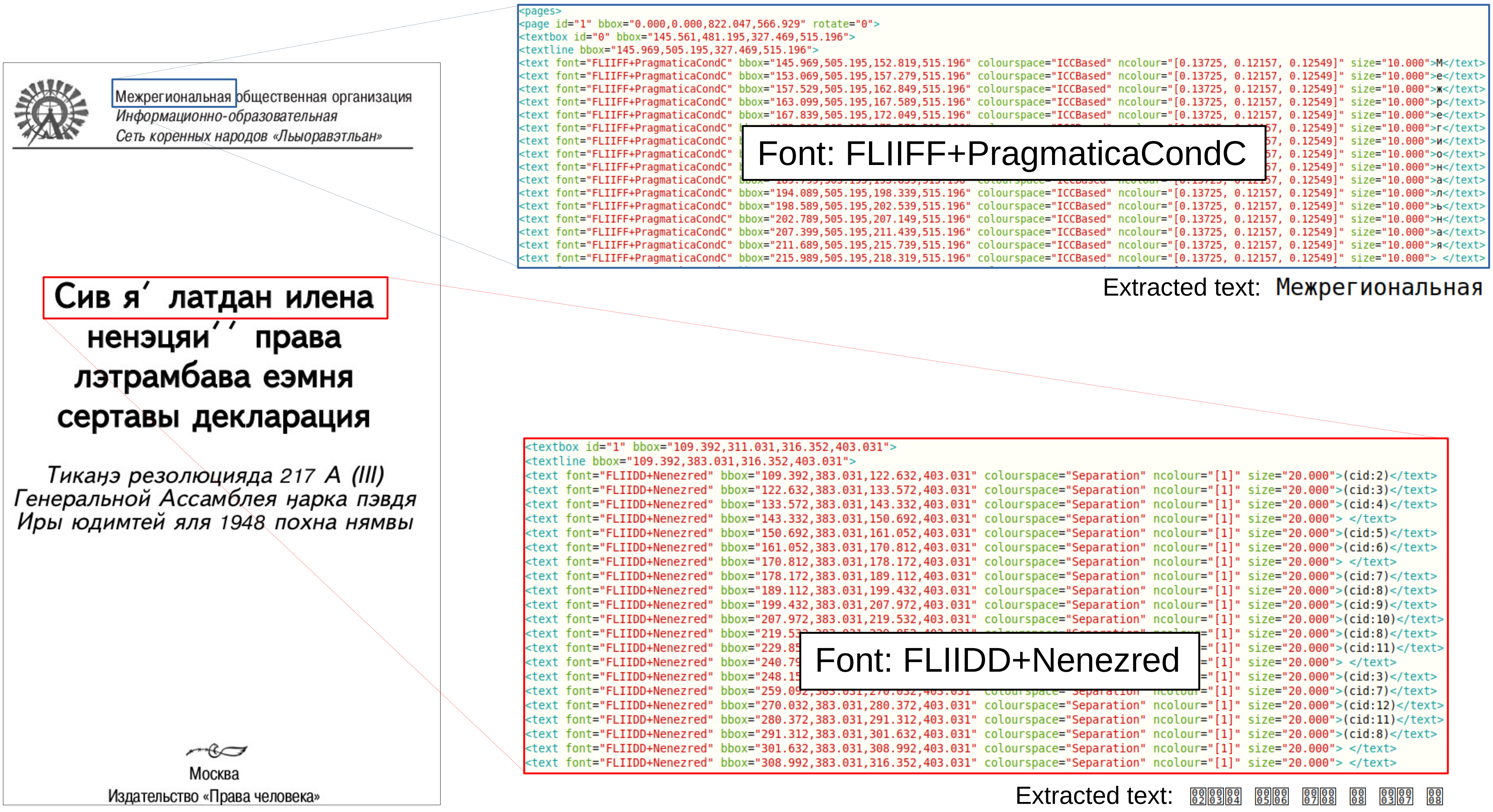
Universal Declaration of Human Rights in Nivkh (niv)

Қатьгун сик правоғун Декларация
Генеральная Ассамблея резолюция 217A (III)
10 к`лолоң 1948 ань жавыл ғера
провозглашайра жад

Нивң қ`атьгун жара, чуғун жара сик намадивңчоғр̆ иввут т`оҳтта, адяй сик п`ңафқ-ңафқ ңазин правоғун ивңы-худ куғытнд жара салғанжунвд жара сикғун мир жара; адяй пиған сикак к`икр̆ п`ид-жаымди қ`оғл ивкғун нивгун правогун лақ қаврта жа-жаңы варвар актғундоҳ вылңута, худғундоҳ нивгун сик макыр қ`оғл легу осқита жагут урла қ`оғл ивң нивгунаҳ п`ңалзайгута; адяй айф тымди мир айғай – ху мирух нивгун кулғытң туғс жара аҳтнд жара ивңы адяй к`лунд жара жекинд жара ағзута – жудғун нивгун язра ныйнынд; адяй нивгун жекирkир̆ п`синңрукғундоҳ озн вопун ухмуна жа ғавргуйныркир̆ нивгун право властьғун торкир̆ ытңуна; адяй қ`атьгун п`ңафқ-ңафқроҳ ңафққалагуйнын ихмына жагуйн; адяй объединенных наций қ`атьгун п`Уставух адяй қ`атьгун правоғун аҳтхымлыта, нивгун п`и намад ивд яймта азмытьғундиң р̆аңғғундиң салғат жумта жатот социальный прогресс жагуйныңа адяй нивгун малғо куғытнд ивт жунвд ургута; адяй государствоғун – членгун Организация объединенных наций ройныт т`оҳтта,

Қ’атьгун сик правоғун Декларация
Генеральная Ассамблея резолюция 217 A (III)
10 қ’лолоң 1948 ань жавыл ғера
провозуглавшайра жад

Нивң қ’атьгун жара, чуғун жара сик намадивңчоғр̆ иввут т’оҳтта, адяй сик п’ңафқ-ңафқ ңазин правоғун ивңы-худ куғытнд жара салғанжунвд жара сикғун мир жара; адяй пиған сикак к’икр̆ п’ид-жаымди қ’оғл ивкғун нивгун правогун лақ қаврта жа-жаңы варвар актғундоҳ вылңута, худғундоҳ нивгун сик макыр қ’оғл легу осқита жагут урла қ’оғл ивң нивгунаҳ п’ңалзайгута; адяй айф тымди мир айғай – ху мирух нивгун кулғытң туғс жара аҳтнд жара ивңы адяй к’лунд жара жекинд жара ағзута – жудғун нивгун язра ныйнынд; адяй нивгун жекирkир̆ п’синңрукғундоҳ озн вопун ухмуна жа ғавргуйныркир̆ нивгун право властьғун торкир̆ ытңуна; адяй қ’атьгун п’ңафқ-ңафқроҳ ңафққалагуйнын ихмына жагуйн; адяй объединенных наций қ’атьгун п’Уставух адяй қ’атьгун правоғун аҳтхымлыта, нивгун п’и намад ивд яймта азмытьғундиң р̆аңғғундиң салғат жумта жатот социальный прогресс жагуйныңа адяй нивгун малғо куғытнд ивт жунвд ургута; адяй государствоғун – членгун Организация объединенных наций ройныт т’оҳтта,


Межрегиональная общественная организация
Информационно-образовательная
Сеть коренных народов «Льыоравэтльан»
Сив я' латдан илена
ненэцяи'' права
лэтрамбава ээмня
сертавы декларация
Тикаӈэ резолюцияда 217 А (III)
Генеральной Ассамблея ӈарка пэвдя
Иры юдимтей яля 1948 похна нямвы
Москва
Издательство «Права человека»
<pages>
<page id="1" bbox="0.000,0.000,822.047,566.929" rotate="0">
<textbox id="0" bbox="145.561,481.195,327.469,515.196">
<textline bbox="145.969,505.195,327.469,515.196">
<text font="FLIIFF+PragmaticaCondC" bbox="145.969,505.195,152.819,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">М</text>
<text font="FLIIFF+PragmaticaCondC" bbox="153.069,505.195,157.279,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">е</text>
<text font="FLIIFF+PragmaticaCondC" bbox="157.529,505.195,162.849,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">ж</text>
<text font="FLIIFF+PragmaticaCondC" bbox="163.099,505.195,167.589,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">р</text>
<text font="FLIIFF+PragmaticaCondC" bbox="167.839,505.195,172.049,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">е</text>
<text font="FLIIFF+PragmaticaCondC" bbox="194.089,505.195,198.339,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">л</text>
<text font="FLIIFF+PragmaticaCondC" bbox="198.589,505.195,202.539,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">ь</text>
<text font="FLIIFF+PragmaticaCondC" bbox="202.789,505.195,207.149,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">н</text>
<text font="FLIIFF+PragmaticaCondC" bbox="207.399,505.195,211.439,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">а</text>
<text font="FLIIFF+PragmaticaCondC" bbox="211.689,505.195,215.739,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000">я</text>
<text font="FLIIFF+PragmaticaCondC" bbox="215.989,505.195,218.319,515.196" colourspace="ICCBased" ncolour="[0.13725, 0.12157, 0.12549]" size="10.000"> </text>
Font: FLIIFF+PragmaticaCondC
Extracted text: Межрегиональная
<textbox id="1" bbox="109.392,311.031,316.352,403.031">
<textline bbox="109.392,383.031,316.352,403.031">
<text font="FLIIDD+Nenezred" bbox="109.392,383.031,122.632,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:2)</text>
<text font="FLIIDD+Nenezred" bbox="122.632,383.031,133.572,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:3)</text>
<text font="FLIIDD+Nenezred" bbox="133.572,383.031,143.332,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:4)</text>
<text font="FLIIDD+Nenezred" bbox="143.332,383.031,150.692,403.031" colourspace="Separation" ncolour="[1]" size="20.000"> </text>
<text font="FLIIDD+Nenezred" bbox="150.692,383.031,161.052,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:5)</text>
<text font="FLIIDD+Nenezred" bbox="161.052,383.031,170.812,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:6)</text>
<text font="FLIIDD+Nenezred" bbox="170.812,383.031,178.172,403.031" colourspace="Separation" ncolour="[1]" size="20.000"> </text>
<text font="FLIIDD+Nenezred" bbox="178.172,383.031,189.112,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:7)</text>
<text font="FLIIDD+Nenezred" bbox="189.112,383.031,199.432,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:8)</text>
<text font="FLIIDD+Nenezred" bbox="199.432,383.031,207.972,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:9)</text>
<text font="FLIIDD+Nenezred" bbox="207.972,383.031,219.532,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:10)</text>
<text font="FLIIDD+Nenezred" bbox="270.032,383.031,280.372,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:12)</text>
<text font="FLIIDD+Nenezred" bbox="280.372,383.031,291.312,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:11)</text>
<text font="FLIIDD+Nenezred" bbox="291.312,383.031,301.632,403.031" colourspace="Separation" ncolour="[1]" size="20.000">(cid:8)</text>
<text font="FLIIDD+Nenezred" bbox="301.632,383.031,308.992,403.031" colourspace="Separation" ncolour="[1]" size="20.000"> </text>
<text font="FLIIDD+Nenezred" bbox="308.992,383.031,316.352,403.031" colourspace="Separation" ncolour="[1]" size="20.000"> </text>
Font: FLIIDD+Nenezred
Extracted text:

1	2	I	—	?	?
Ӈ	ӈ	I	—	ё	А
?	?	?	?	?	?
Н	О	П	С	Т	У
?	?	?	?	?	?
д	е	ж	з	и	й
.	?	?	1	2	?
с	т	у	ф	х	ц
?	?				
ю	я				

Corrupted Nenets (yrk) font

d	;	t	ê	H
d	;	t	ê	H
N	ú	E	Q	O
N	ú	E	Q	O
A	c	G	I	-
A	c	G	I	-
.	ç	h	1	ã
.	ç	h	1	ã

Valid Portuguese (por) font

Recovery process

```
...
line 90 => 1 10 6 7 4 3 17
line 91 => 3 6 5 6 5 4
line 92 => 1 10 3 4 8 2 3 11 2 4
line 93 => 4 3 4 9 6 6 5 4
line 94 => 6
line 95 => 9 5 4
line 96 => 1 11 3 3 11 9 5 6
line 97 => 1 3 3 3 12 2 4 12 2 4
line 98 => 11 12 5 10 11
line 99 => 2 5 2 8 7 5 2 6
line 100 => 1 10 14 5 4
line 101 => 1 11 3 6 8 2 11 5
line 102 => 5 4 2 11 11 5 3 5 5
...
```

Unicode sequence (input from user):

Организация Объединенных Наций цельгундох ёскиндфурнд

→ token lengths: [11, 12, 5, 10, 11]

CID sequence (extracted from the document) + automatic guess of space and dot characters:

:24:48:35:32:45:40:39:32:54:40:63 :24:33:58:37:36:40:45:37:45 :45:59:53 :23:32:54:40:41 :54:37:43:60:8:51:45:36:46:9 :16:49:42:40 :45:36:52:51:48:45:36

→ token lengths: [11, 12, 5, 10, 11]

Unique match!

```
new 24 -> O
new 48 -> p
new 35 -> г
new 32 -> а
new 45 -> н
new 40 -> и
new 39 -> з
…
```

28 symbols recovered → 34% of the characters

l.0001: :23:40:34:6 :10:15:32:50:60:35:51:45 :5:32:48:32:2 :55:51:8:51:45 :5:32:48:32 :49:40:42 :45:32:44:32:36:40:34:6:55:46:8:7 :40:34:34:51:50
l.0002: :50:15:46:9:50:50:32:2 :32:36:63:41 :49:40:42 :47:15:6:32:52:10:3:6:32:52:10 :6:32:39:40:45 :47:48:32:34:46:8:51:45 :40:34:6:59:53:51:36 :42:51:8:59:50:45:36
l.0003: :5:32:48:32 :49:32:43:12:32:45:5:51:45:34:36 :5:32:48:32 :49:40:42:8:51:45 :44:40:48 :5:32:48:32 :32:36:63:41 :47:40:12:32:45 :49:40:42:32:42 :42:15:40:42:7
l.0004: :47:15:40:36:3:5:32:59:44:36:40 :10:15:46:12:43 :40:34:42:8:51:45 :45:40:34:35:51:45 :47:48:32:34:46:35:51:45 :43:32:10 :10:32:34:48:50:32 :5:32:3:5:32:6:59
l.0005: :34:32:48:34:32:48 :32:42:50:8:51:45:36:46:9 :34:59:43:6:51:50:32:2 :53:51:36:8:51:45:36:46:9 :45:40:34:35:51:45 :49:40:42 :44:32:42:59:48 :10:15:46:12:43

l.0001: Ни:34:6 :10:15:а:50:ьгун :5:ара:2 :55:уғун :5:ара сик на:44:ади:34:6:55:оғ:7 и:34:34:у:50
l.0002: :50:15:оӂ:50:50:а:2 адяй сик :47:15:6:аф:10:3:6:аф:10 :6:азин :47:ра:34:оғун и:34:6:ыхуд куғы:50:нд
l.0003: :5:ара сал:12:ан:5:ун:34:д :5:ара сикғун :44:ир :5:ара адяй :47:и:12:ан сикак к:15:ик:7
l.0004: :47:15:ид:3:5:аы:44:ди :10:15:о:12:л и:34:кғун ни:34:гун :47:ра:34:огун ла:10 :10:а:34:р:50:а :5:а:3:5:а:6:ы
l.0005: :34:ар:34:ар ак:50:ғундоӂ :34:ыл:6:у:50:а:2 худғундоӂ ни:34:гун сик :44:акыр :10:15:о:12:л

Нивң қ`атьгун ӂара, чуғун ӂара сик намадивңчоғр̆ иввут т`оҳтта, адяй сик п`ңафқ-ңафқ ңазин правоғун ивңы-худ куғытнд ӂара салғанӂунвд ӂара сикғун мир ӂара; адяй пиған сикак к`икр̆ п`ид-ӂаымди қ`оғл ивкғун нивгун правогун лақ қаврта ӂа-ӂаңы варвар актғундоҳ вылңута, худғундоҳ нивгун сик макыр қ`оғл

221=>жымди қ`оғл уйгид
Ниғвң дуфтож выл̧уд Санги
Нивң қ`атьгун жара, чуғун жара сик намадивң̧чоғр̧
Организация Объединенных Наций цельғундож ёскиндфурнд
эна положенияяғун ивғай напы п`жатьжать қаврна, п`ң̧афқ-ң̧афқ
самоуправляющаяся жа ғаврд лу,
Генеральная Ассамблея туң̧ сик
Декларация задача жагун провозглашайдра
Чу п`жоғара, сикак маң̧ра жаң̧ общество
жекинд
удовлетворить
зақоўид.
Произвольно
Техническое
п`Уставух
Конституцияғир̧
Ин
Қ`атьң̧гун
Эғлгун
Ыткғун
Р̧аң̧гымкмунд

…
line 175 => 7 6 5 8 4 2 6 3 4
line 176 => 8 11 2 4 7 5 5 6
line 177 => 10 6 6 5 4
…
line 220 => 12 7 8 4 10 4 7
line 221 => **5 5 6**
line 222 => 7 5 12 4 1 15
…

Automatically guesses characters for space and dot

Sorts characters by number of instances + displays sample lines containing them:

```
* cid 25 count: 2
line 101   :25:роиз:34:ольно :5:а:50 на:7:50:и:6 нарух:50:и:6 йа гра:38:данс:50:34:о ас:10:а:44
line 209   :25:15:и :47:ра:34:оғун :5:ара куғы:50:ндғун :5:ара осу:57:ес:50:34:ляй:6:ы ни:34:гун
```

Sorts lines, by number of unrecovered characters + display sample:

```
line 172 ['6', '44', '5', '5'] рогуйны:6 :44:едицинское :5:ара социальное :5:ара иғрығрынд
line 34 ['44', '6', '6', '34'] леле:44:а:6:6 образо:34:ание
```

Checks consistency of user input

Automatically guesses upper case letters based on document statistics

```
GUESS CAPS
try: :1:аңы
MATCH! жаңы
new 1 -> Ж
```

n	1	2	3	4	5	6
niv	10	33	72	93	97	98
yrk	17	29	60	86	98	100

Table 1: Proportion in percent of unique sequences of token lengths for sequence length n in the UDHR for two different languages

language	niv	yrk
unique characters	81	68
text length (words)	1430	1530
input length (words)	57	76

Table 2: Unique character count and total word length for the UDHR declaration in two languages, and the total number of input words necessary for full recovery of these texts

Discusion, conclusion, questions...

See you at the poster presentation! :-)