Име овог портала је изабрано у знак сећања на програмски систем АУРОРА (Витас, 1979). који представља један од првих корака у аутоматској обради писаних текстова на српском језику. Овде ће корисник наћи део оних информација какве је некада генерисала АУРОРА, а које су данас добијене применом корпусног процесора Unitex/GramLab[1], програмског система у који су интегрисане бројне полазне идеје о обради улазног текста на српском.

Сврха портала је да пружи истраживачима српске књижевности, али и другим заинтересованим корисницима, „микроскопски“ увид у лексику појединих дела српске књижевности нудећи, уз сам текст, и његове конкорданце и фреквенцијар, као и навигацију између текста и списка речи које су из њега ексцерпиране.

Назив АУРОРА је акроним од *АУтоматска Рутина за Обраду РечникА*. Овај програм је био написан у крајње неподесном језику за обраду текста, у језику Fortran 66, а извршавао се на машини IBM 360/44[2] у ондашњем Рачунском центру Математичког института у Београду.

Подстицај за конструкцију АУРОРЕ је дошао са првог скупа о обради језичких података који је у децембру 1977, организовао Милан Шипка у Сарајеву (Шипка, 1978). На овом скупу су словеначки и хрватски информатичари и лингвисти приказали своје радове са подручја обраде корпуса, а посебно програме за израду конкорданци текста. Програмски систем који се користио за ове намене била је у то доба COCOA (Коркоран, 1974). Како овај програм није био доступан за машину којом смо располагали, није било другог пута до да се напише властити систем.

Обрада у Аурори је била замишљена тако да не зависи ни од језика, ни од писма на коме је текст написан. То је значило да се не може подразумевати колациона секвенција индукована кодом машине у уазбучавању текста, као и да сама дефиниција појма речи мора бити флексибилна. Како је IBM 360/44 користио као носач за улазне податке 80-колонску бушену картицу, репертоар карактера, у коме се улазни текст могао кодирати, је био ограничен на велика слова енглеске абецеде, цифре декадног ситема и известан број посебних карактера (аритметички знаци и симбол $). Под таквим ограничењима је настала кодна шема која се и данас назива АУРОРА, а која се користи приликом постављања упита на корпусу савременог српског језика[3]. Незнатно измењена верзија овог кода се налази и у Правопису (Пешикан, 1993) под називом *телепринтерски код*. Идеја оваквог кодирања је била да се каракарктери који се не користе у текстовима на српском (као *X, Y, Q, W*) употребе у пресликавању српског алфабета на репертоар расположивих карактера на бушеној картици. На овај начин је била неутралисана разлика између ћириличног и латиничног писма у обради корпуса. Неупотребљени карактери су били резервисани за један скромни опис изгледа текста. Тако је симбол $ био коришћен за обележавање наслова у тексту, + да означи велико слово, а / W крај реда. Овакав прилаз је омогућио да се и појам речи и лексикографски поредак дефинишу у зависности од потреба обраде. Пример текста кодираног на овај начин је дат на слици 1.

```
$5 2. CYOVEK I PRIRODA $ W
+PRE VISXE HILXADA/ W
GODINA CYOVEK JE BIO BESPOMOCXNO BICXE PRED/ W
PRIRODOM KOJA GA JE OKRUZXAVALA. +TESXKO JE/ W
NABAVLXAO HRANU. +S NAPOROM SE ZASXTICXAVAO/ W
OD HLADNOCXE I VRUCXINE. +KORISTECXI SE ISKU-/ W
STVOM, IZGRADXIVAO JE ZA SVOJE POTREBE PROSTA/ W
ORUDXA. +OTKRIVAO JE OSOBINE ZXIVOTINXA, BI-/ W
LXAKA I MRTVE PRIRODE. +TAKO SU LXUDI, KORAK/ W
PO KORAK, SAVLADXIVALI PRIRODU I POSTAJALI/ W
NXENI GOSPODARI./ W
$5 3. ODNOSI MEDXU LXUDIMA $ W
```

Слика 1. Извод из *Уџбеника историје* за 5. разред основне школе (Божић. 1982)

И појам речи је било могуће флексибилно дефинисати. Тако су, примера ради, када су у скуп сепаратора додати вокали, као „речи“ у конкорданцама приказиване консонантске групе у српском (Крстев, 1985).

Ток обраде текста се састојао у конструкцији инвертованог текста: индекса који упућује на оланчану листу свих појављивања различитих речи, а за сваку речу - на листу позиција њених појављивања. Листа појављивања различитих речи је била вишеструко оланчана (по почетном и

[1] https://unitexgramlab.org/

[2] Овај рачунар се данас налази у београдском Музеју науке и технике.

[3] http://www.korpus.matf.bg.ac.rs/prezentacija/uputstvo.html

завршном слову, по фреквенцији, итд). Оваква интерна репрезентација текста омогућавала је, с једне стране, да се из ње директно конструишу конкорданце према различитим критеријумима. С друге стране, повезивањем интерне репрезентације текста са другим програмима је пружило могућност различитих експеримената. Тако је повезивање АУРОРЕ и морфолошког генератора именских речи (Витас, 1980) омогућило експерименте у претрази корпуса према именској одредници (леми), а не према њеним појединачним облицима.

Прве примене, које су истовремено биле и почетак прикупљања текстова за корпус савременог српског језика, су биле везане за сарадњу са Заводом за уџбенике и наставна средства на пројекту изучавања развоја ученичког говора. Овај пројекат је обухватио формирање корпуса од опсежних одговора ученика одређеног узраста на унапред утврђена питања. Лексика овако добијеног корпуса је затим упоређивана са лексиком одговарајућих уџбеника (Цвијовић, 1984). Друге занимљиве примене у оно доба су биле везане за анализу лексичког богатства језика затвореника (Пантазијеић, 1985) или за анализу језика закона. Верзије Ауроре су дале и прве експерименте у откривању типографских грешака (Витас, 1985), прву обраду параленог корпуса (Крстев, 1988) кроз поређење корпуса упутстава за лекове на српском и словеначком или основу за израду индекса за превод *Оксфордског речника рачунарства* (Illingworth, 1990) и *Вукових пословица* (Вук, 1996), (Крстев, 1997).

Ипак, полазни мотив је у оно време била обрада књижевних текстова. Требало је да прође скоро пола века како би технологија омогућила да анализе какве је генерисала АУРОРА буду доступне најширем кругу корисника. У овом облику, АУРОРА пружа увид у речник појединих књижевних дела и представља почетни материјал за израду речника појединих писаца. Будуће верзије овог портала, које би користиле пуни садржаја система електронских морфолошких речника за српски језик, ће давати и финији увид у књижевно дело. Поменимо овде, као пример праваце будућег развоја, лематизирање конкорданци или опремање речи у индексу семантичким атрибутима које садрже електронски речници.

(Божић, 1982) Божић, Иван. *Историја за V разред основне школе*. Завод за уџбенике и наставна средства, Београд

(Витас, 1979) Витас, Душко. Приказ једног система за аутоматску обраду текста, *INFORMATICA'79,* Блед, стр. 7 101

(Витас, 1980) Витас, Душко: Генерисање именичких облика у српскохрватском. *Informatica 80(3),* Словеначко друштво за информатику, Љубљана, 49-55

(Витас, 1985) Витас, Душко: Један поступак аутоматске сегментације српскохрватских речи и његове примене, Зборник са III научног скупа "*Рачунарска обрада језичких података*", Институт Јожеф Стефан, Блед, 3–14

(Вук, 1996) Караџић, Вук Стефановић: *Вукове народне пословице с регистром кључних речи*. Београд : Нолит

(Illingworth, 1990) Illingworth, Valerie: *Oksfordski rečnik računarstva*, Београд: Нолит

(Krstev, 1985) Krstev, Cvetana: Rastavljanje reči srpskohrvatskog jezika na kraju retka", u Zbornik radova sa III naučnog skupa "Računarska obrada jezičkih podataka", Institut "Jožef Štefan" Ljubljana, 289-301

(Krstev, 1988) Krstev, Cvetana; Smiljana Jović-Puač i Duško Vitas: Analiza podjezika uputstava za lekove na srpskohrvatskom i slovenačkom jeziku, u Zbornik radova sa IV naučnog skupa "*Računarska obrada jezičkih podataka*", Institut "Jožef Štefan", Ljubljana, 249-255

(Крстев, 1997) Krstev, Cvetana: Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije, докторска дисертација, Математички факултет, Универзитет у Београду

(Коркоран, 1974) Paul E. Corcoran. "COCOA: A FORTRAN Program for Concordance and Word-count Processing of Natural Language Texts". *Behavior Research Methods & Instrumentation*. 6 (6): 566. doi:10.3758/BF03201351

(Pantazijević, 1985) Pantazijević-Stanojević, Milica: *Čovek lišen slobode i literatura: resocijalizacija*, Viša škola unutrašnjih poslova, Beograd

(Пешикан, 1993) Пешикан, Митар ; Јерковић, Јован ; Пижурица, Мато: *Правопис српскога језика*. Матица српска. Нови Сад

(Cvijović, 1984) Cvijović, Milka: *Dečji govor : rečnik i rečenica : četvrti razred osnovne škole*, Zavod za udžbenike i nastavna sredstva, Beograd

(Шипка, 1978) Шипка, Милан (уредник). *Kompjuterska obrada lingvističkih podataka*. Posebna izdanja 4, Institut za književnost i jezik, Sarajevo