http://scholarpage.org/ouamour USTHB University

*Important Note : The dataset is totally free for academic and research purposes.

You can obtain your Copy by simply contacting *Dr Siham Ouamour* at the following email address: siham.ouamour@uni.de

DATASET for LANGUAGE IDENTIFICATION (DLI32 and DLI32-2) with 32 different Languages*

* Authors: K. Abainia, S. Ouamour and H. Sayoud

*

* Years: 2013-2015

*

* LANGUAGES	French, English, Arabic, Russian, German, Italian,
*	Greek, Spanish, Persian, Chinese, Turkish, Finnish,
*	Portuguese, Roman, Polish, Hungarian, Dutch, Irish,
*	Icelandic, Hindi, Czech, Malay, Bulgarian, Norwegian,
*	Urdu, Thai, Indonesian, Danish, Hebrew, Swedish, Latin,
*	and Albanian.

DESCRIPTION

DLI32 and DLI32-2 are two corpus used in automatic language identification of written texts. They are collected from different discussion forums, and contain noisy texts encoded with UTF-8 encoding. The texts may contain any kind of the following noises: URLs, Citations in other language, Tags, Abbreviations, Unaccented characters, Typing errors, Html tags and objects, Insignificant characters and SMS writing style.

The DLI32 corpus contains 320 text, which correspond to 10 texts per language. The text size ranges between 93 and 146 words / text, whereas DLI32-2 contains 640 short texts, which correspond to 20 texts per language. The text size ranges between 43 and 67 words / text.

****

****

*** DLI32 dataset content (320 texts):**

****

French: 1.txt - 10.txt

English: 11.txt - 20.txt

Arabic: 21.txt - 30.txt

Russian: 31.txt - 40.txt

German: 41.txt - 50.txt

Italian: 51.Txt - 60.txt

Greek: 61.txt - 70.txt

Spanish: 71.txt - 80.txt

Persian: 81.txt - 90.txt

Chinese: 91.txt - 100.txt

Turkish: 101.txt - 110.txt

Finnish: 111.txt - 120.txt

Hebrew: 121.txt - 130.txt

Portuguese: 131.txt - 140.txt

Roman: 141.txt - 150.txt

Polish: 151.txt - 160.txt

Hungarian: 161.txt - 170.txt

Dutch: 171.txt - 180.txt

Irish: 181.txt - 190.txt

Swedish: 191.txt - 200.txt

Latin: 201.txt - 210.txt

Icelandic: 211.txt - 220.txt

Hindi: 221.txt - 230.txt

Czech: 231.txt - 240.txt

Malay: 241.txt - 250.txt

Bulgarian: 251.txt - 260.txt

Norwegian: 261.txt - 270.txt

Albanian: 271.txt - 280.txt

Urdu: 281.txt - 290.txt

Thai: 291.txt - 300.txt

Indonesian: 301.txt - 310.txt

Danish: 311.txt - 320.txt

*** DLI32-2 dataset content (640 texts):**

****

French: 1.txt - 20.txt

English: 21.txt - 40.txt

Arabic: 41.txt - 60.txt

Russian: 61.txt - 80.txt

German: 81.txt - 100.txt

Italian: 101.Txt - 120.txt

Greek: 121.txt - 140.txt

Spanish: 141.txt - 160.txt

Persian: 161.txt - 180.txt

Chinese: 181.txt - 200.txt

Turkish: 201.txt - 220.txt

Finnish: 221.txt - 240.txt

Hebrew: 241.txt - 260.txt

Portuguese: 261.txt - 280.txt

Roman: 281.txt - 300.txt

Polish: 301.txt - 320.txt

Hungarian: 321.txt - 340.txt

Dutch: 341.txt - 360.txt

Irish: 361.txt - 380.txt

Swedish: 381.txt - 400.txt

Latin: 401.txt - 420.txt

Icelandic: 421.txt - 440.txt

Hindi: 441.txt - 460.txt

Czech: 461.txt - 480.txt

Malay: 481.txt - 500.txt

Bulgarian: 301.txt - 520.txt

Norwegian: 321.txt - 540.txt

Albanian: 341.txt - 560.txt

Urdu: 361.txt - 580.txt

Thai: 381.txt - 600.txt

Indonesian: 401.txt - 620.txt

Danish: 421.txt - 640.txt