TSNE

Unigram

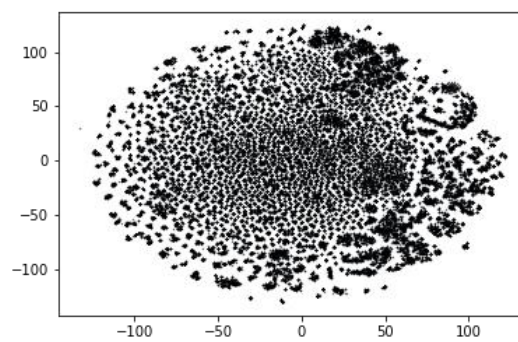**(1) Amharic**

(2) Arabic

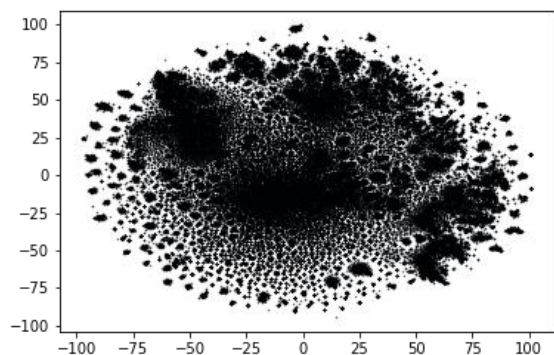
(3) Atikamekw
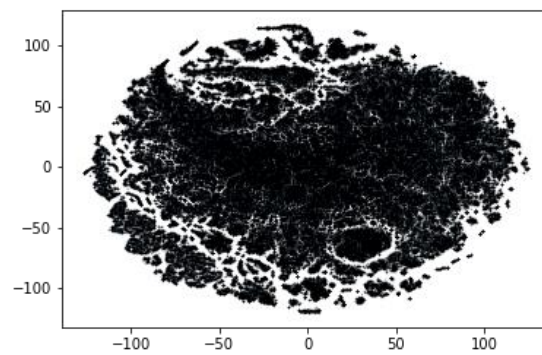

(4) Basque

(5) Belarusian

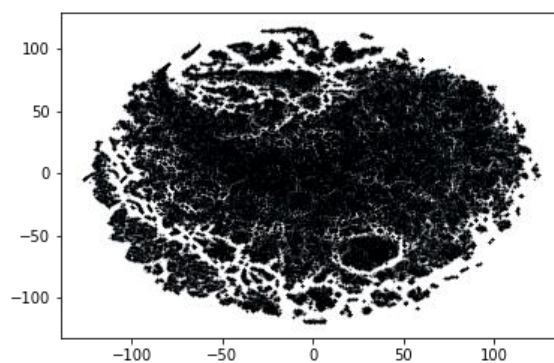

(6) Bengali

(7) Bulgarian
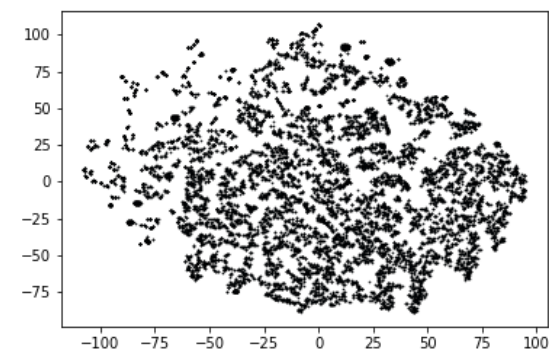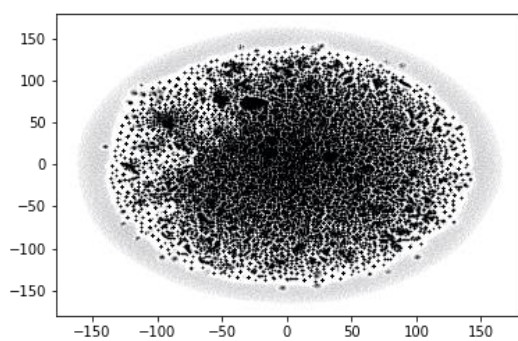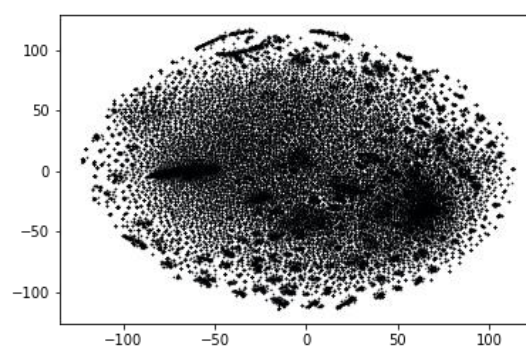

(8) Chechen

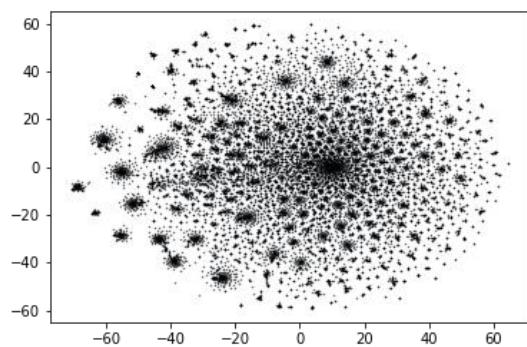**(9) Chinese**

(10) Coptic
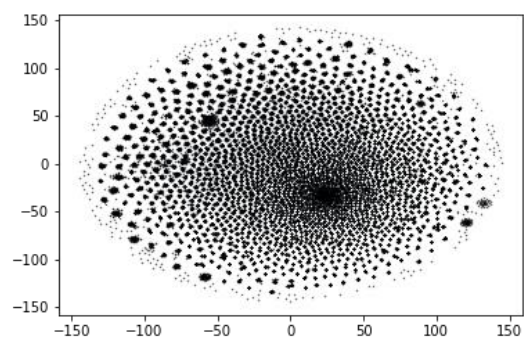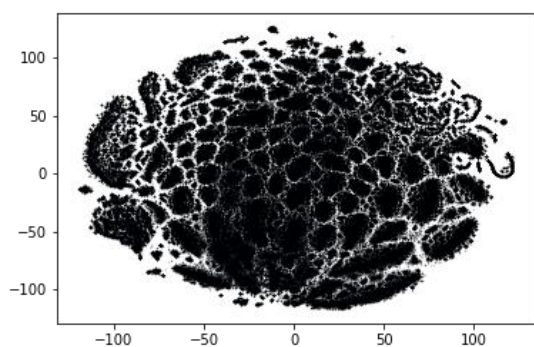

(11) Czech
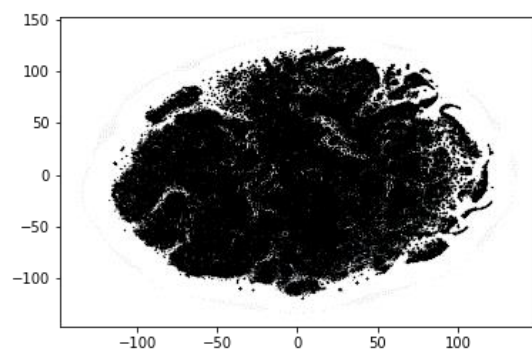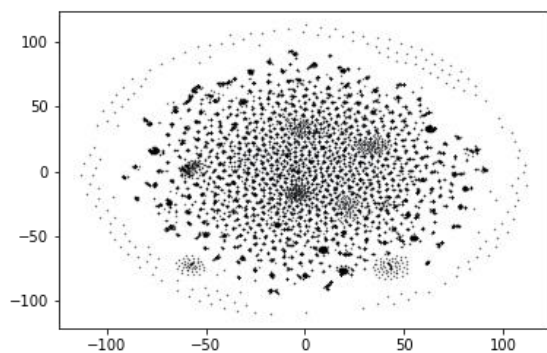

(12) Dholuo

(13) Dutch

(14) English

(15) Esperanto

(16) Finnish

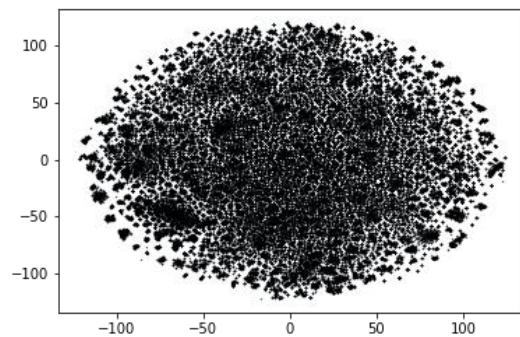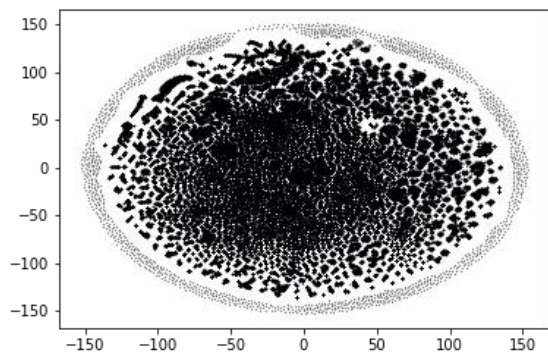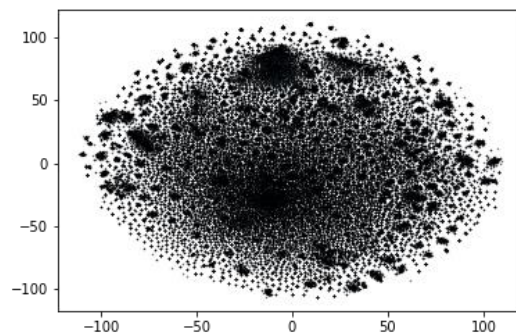

(17) French

(18) German

(19) Hindi
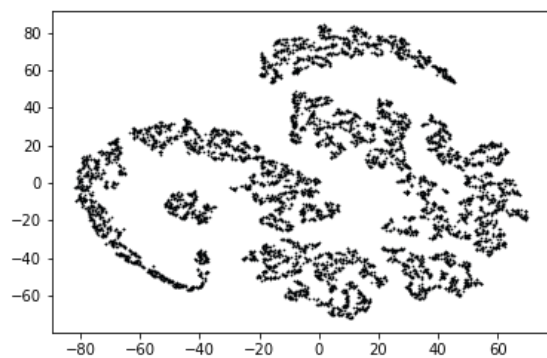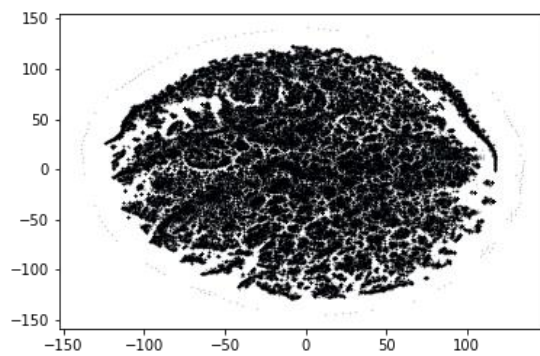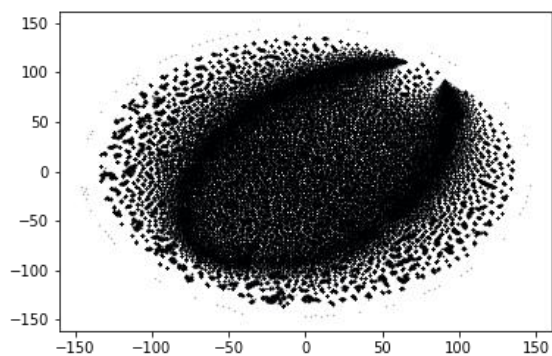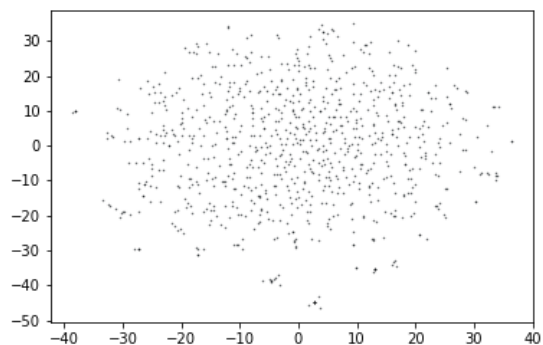

(20) Icelandic

(21) Indonesian

(22) Japanese
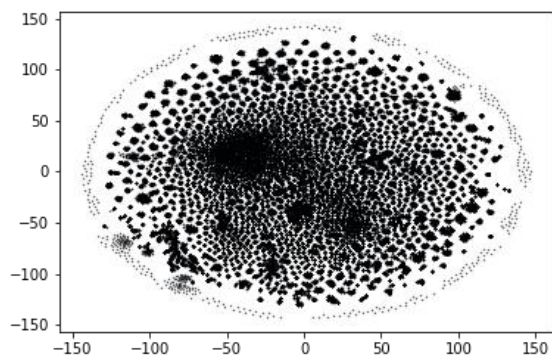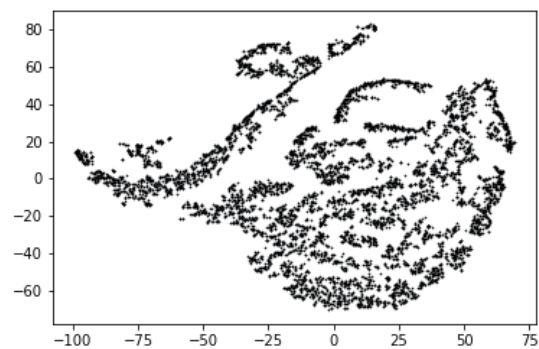
(23) Kabyle

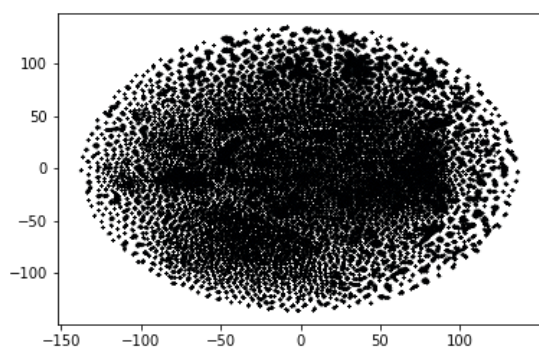**(24) Kazakh**

(25) Latin

(26) Hungarian
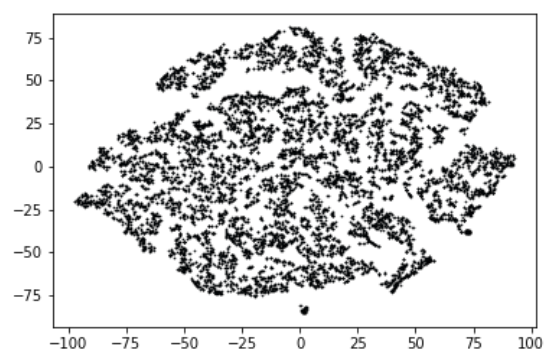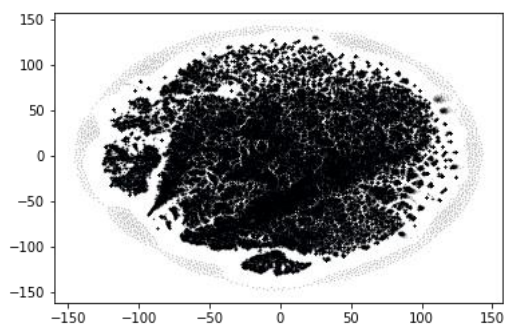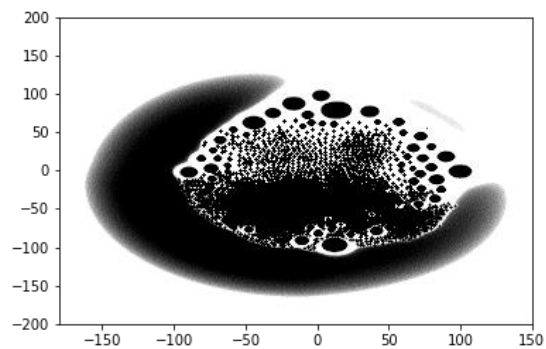

(27) Malayalam

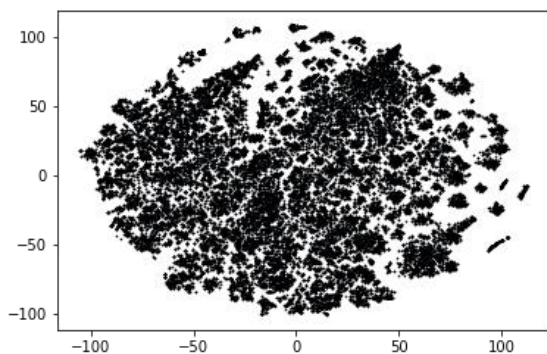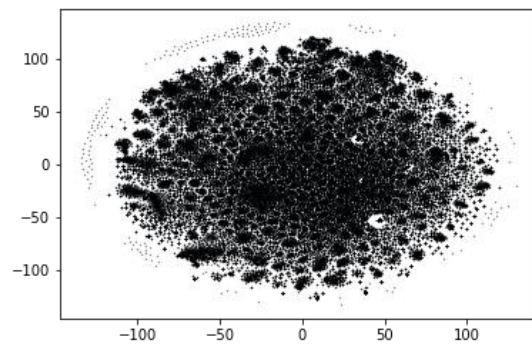**(28) Navajo**

(29) Norwegian

(30) Oromo

(31) Ossetian
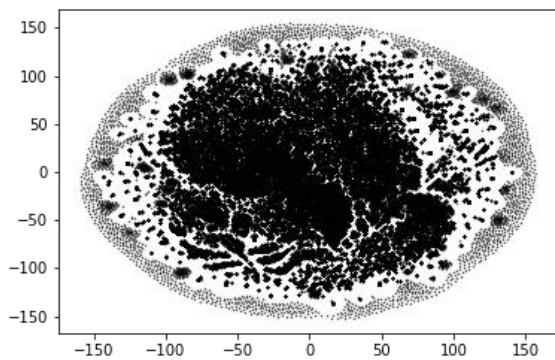

(32) Persian

(33) Polish

(34) Punjabi

(35) Quechua

(36) Romanian

(37) Russian

(38) Serbian

(39) Sinhala

(40) Spanish

(41) Swahili
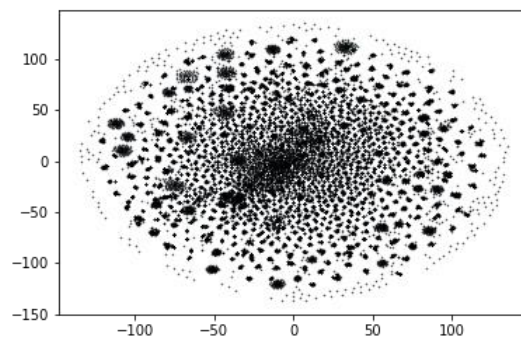

(42) Swedish

(43) Tabasaran

(44) Tagalig
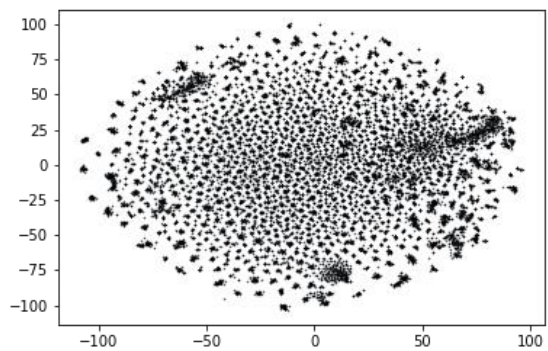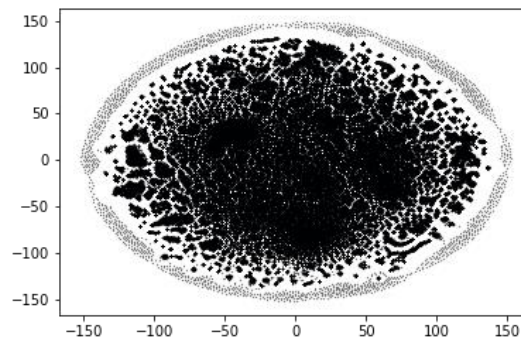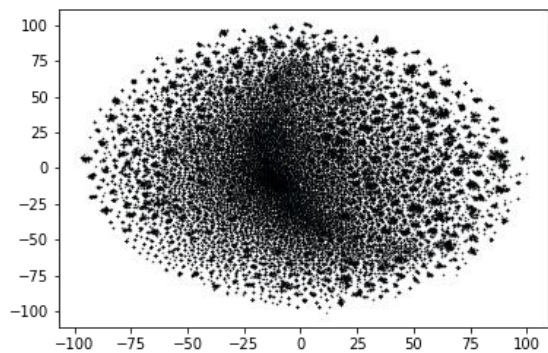

(45) Tatar

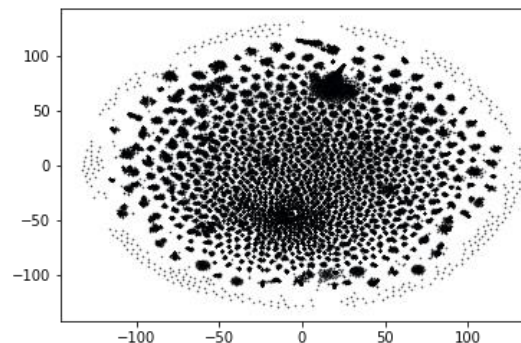**(46) Thai**

(47) Turkish

(48) Tuvan

(49) Udmurt

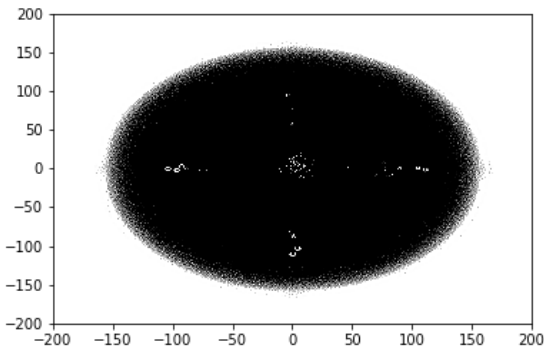**(50) Ukrainian**
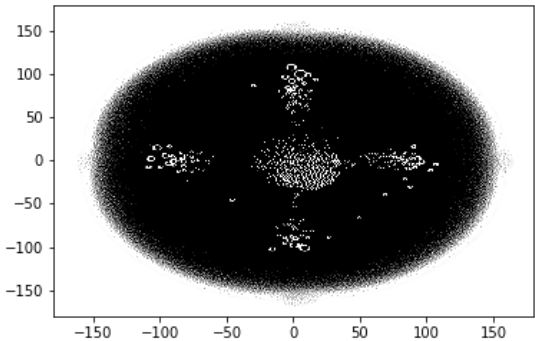
(51) Uzbek

(52) Vietnamese

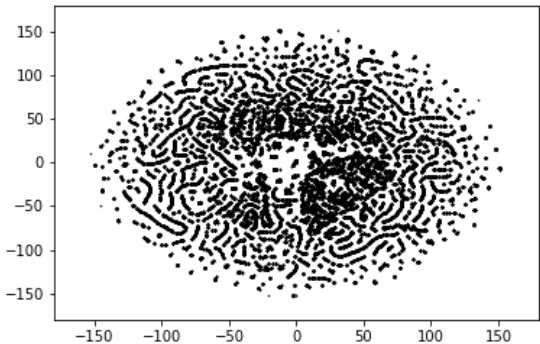Fig. 8: t-SNE visualization of Unigram for 52 languages.
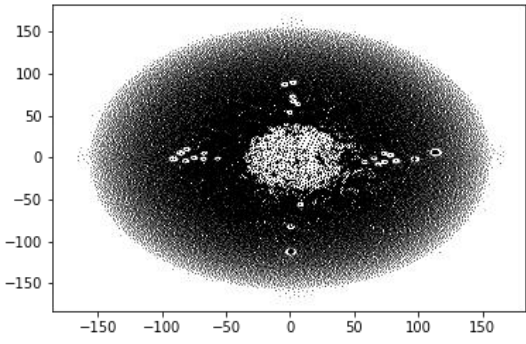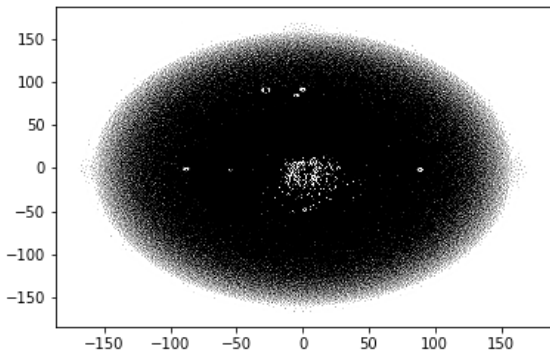
TSNE
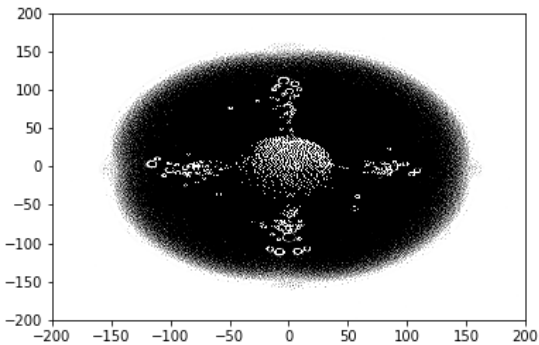
Bigrams
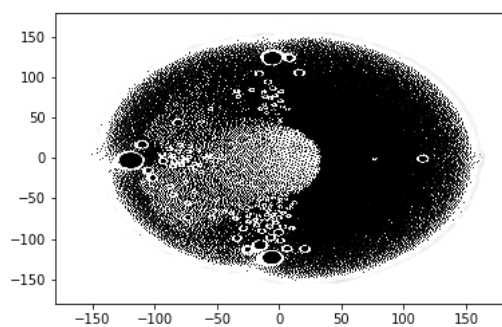

(1) Amharic

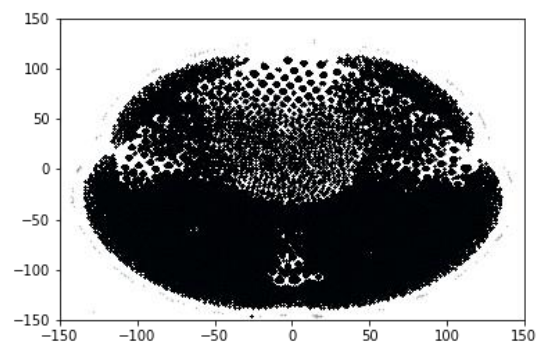**(2) Arabic**

(3) Atikamekw
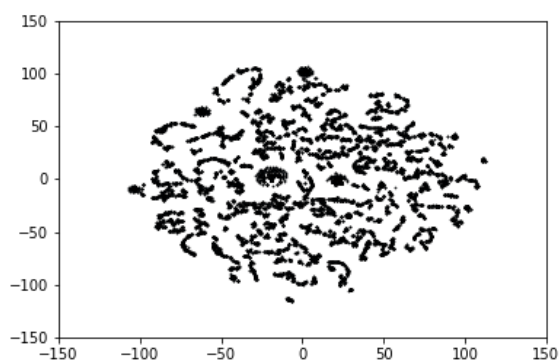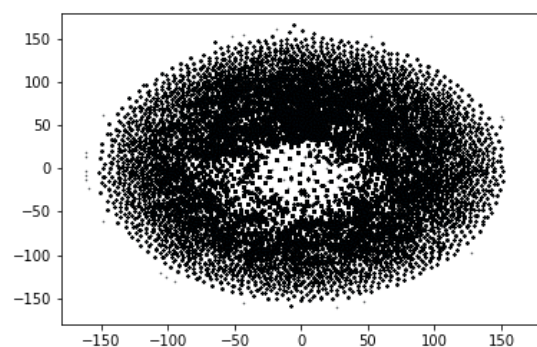

(4) Basque

(5) Belarusian

(6) Bengali

(7) Bulgarian
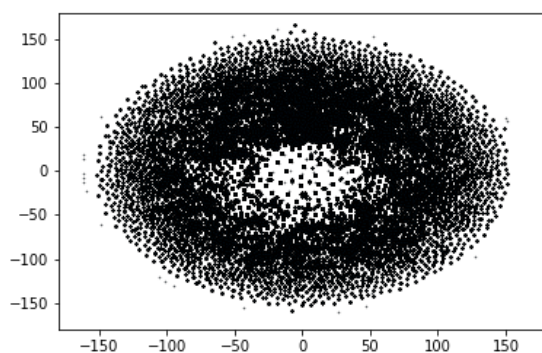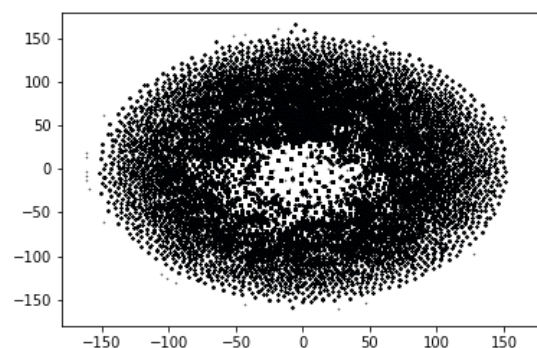

(8) Chechen

(9) Chinese

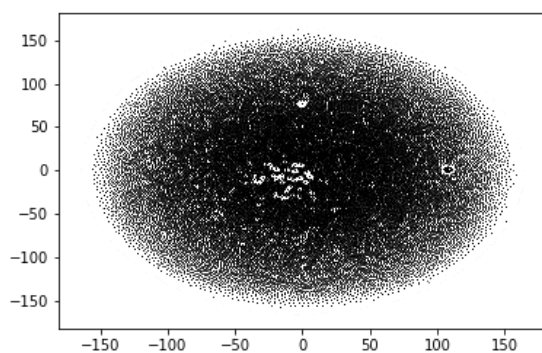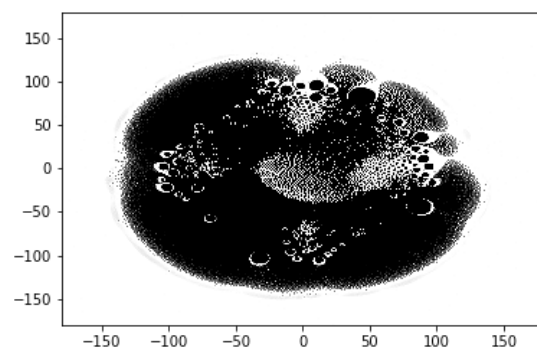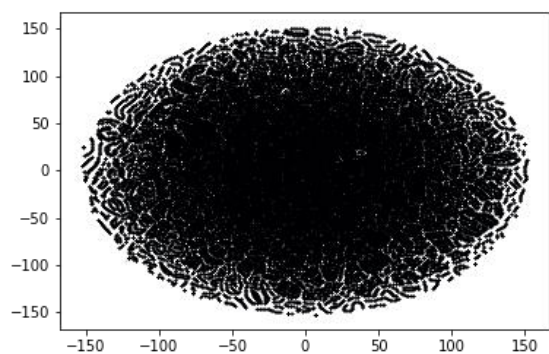

(10) Coptic

(11) Czech

(12) Dholuo
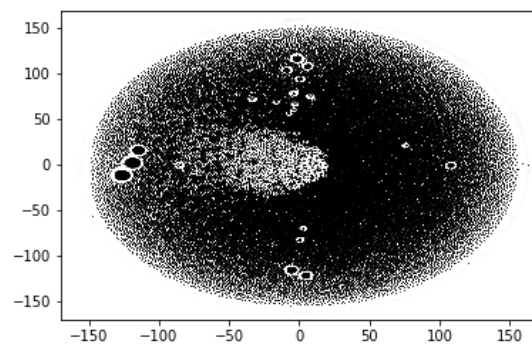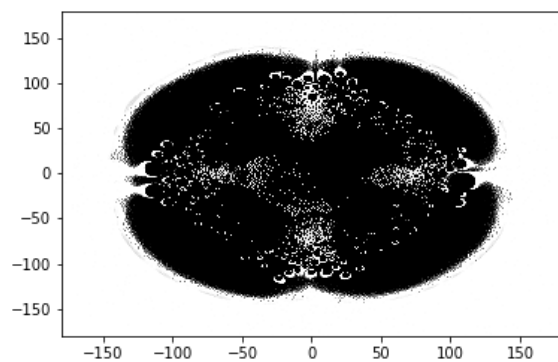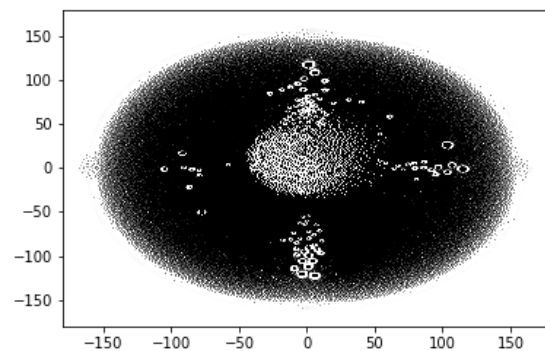

(13) Dutch

(14) English
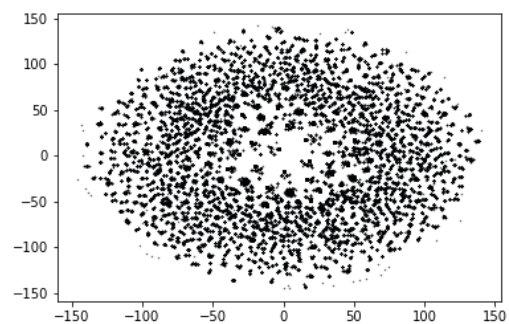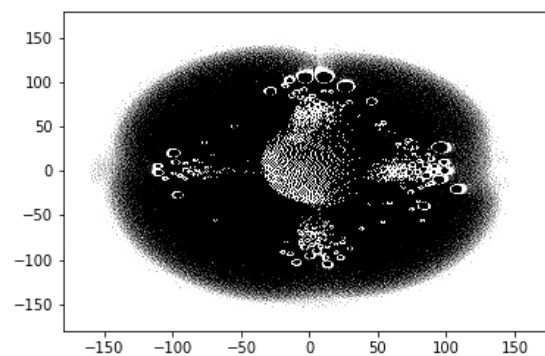
(15) Esperanto

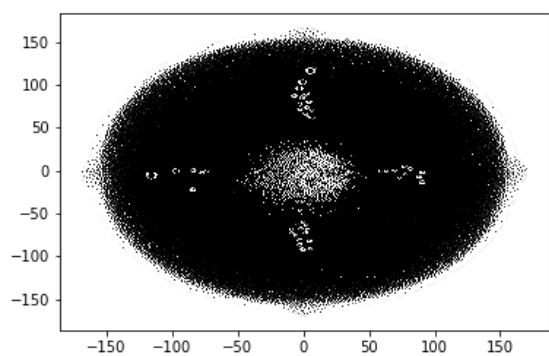
(16) Finnish

(17) French
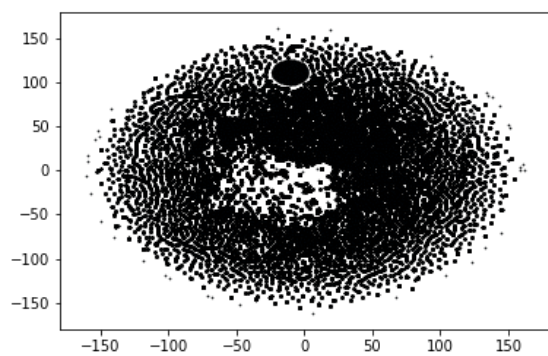

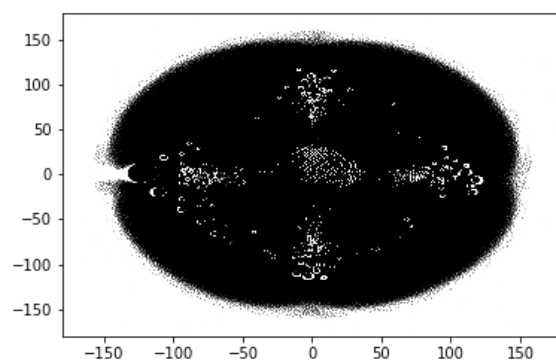**(18) German**

(19) Hindi
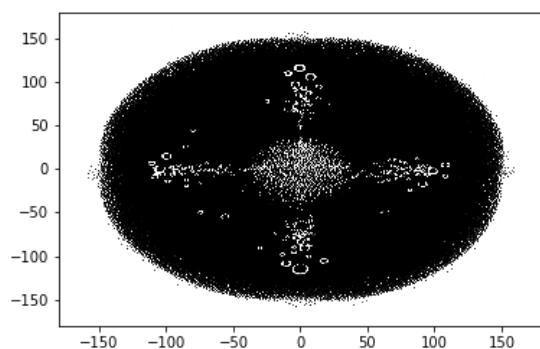

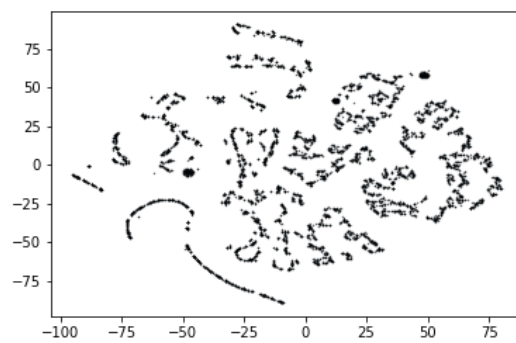**(20) Icelandic**

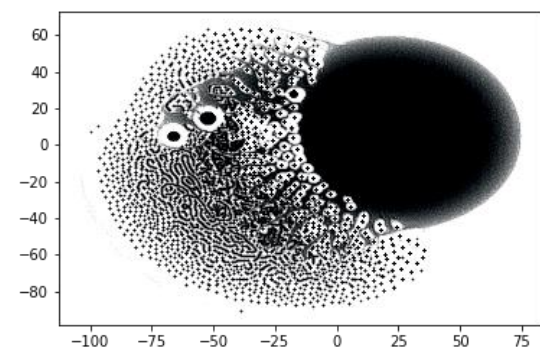**(21) Indonesian**

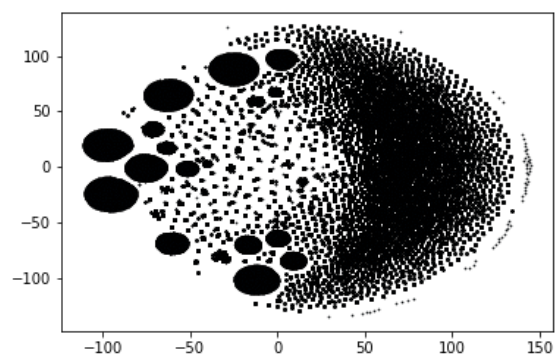**(22) Japanese**

(23) Kabyle

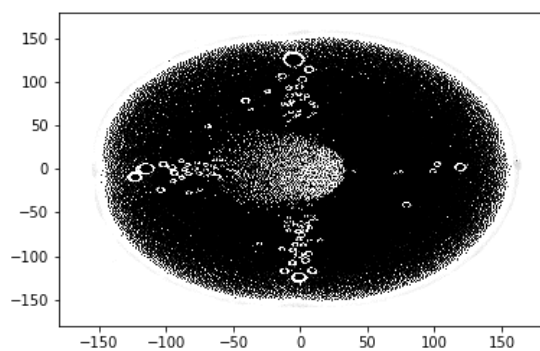**(24) Kazakh**

(25) Latin

(26) Hungarian
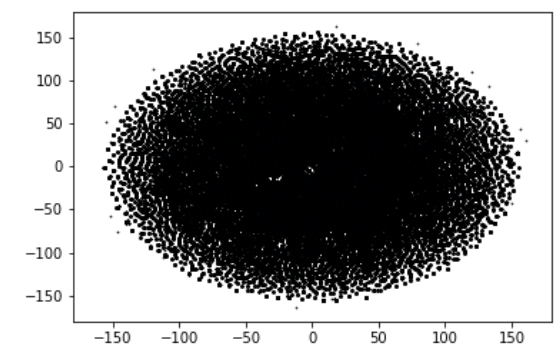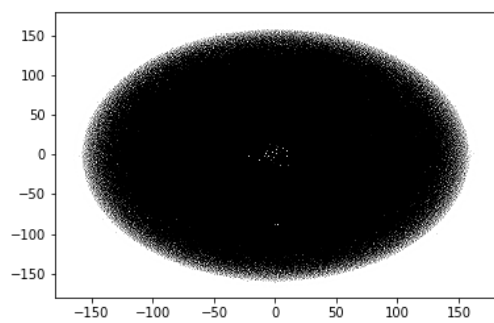

(27) Malayalam

(28) Navajo

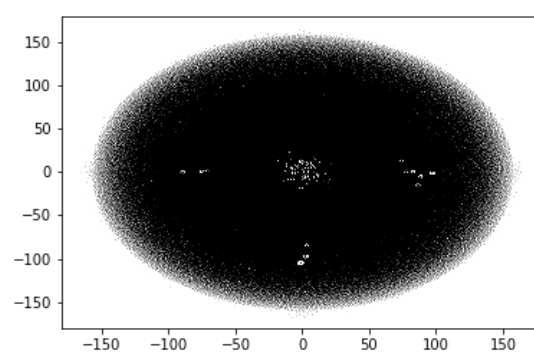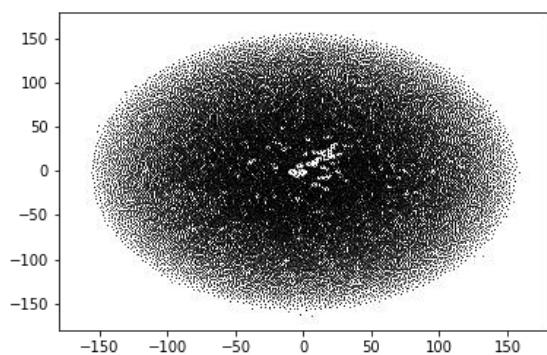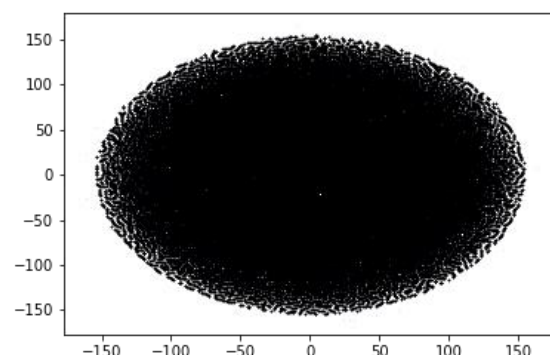
(29) Norwegian

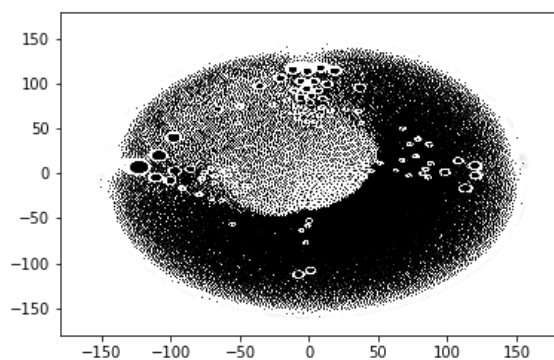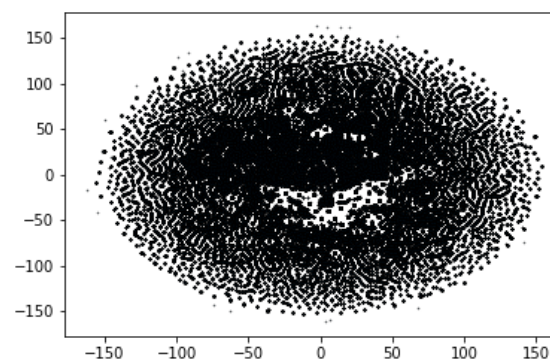**(30) Oromo**

(31) Ossetian

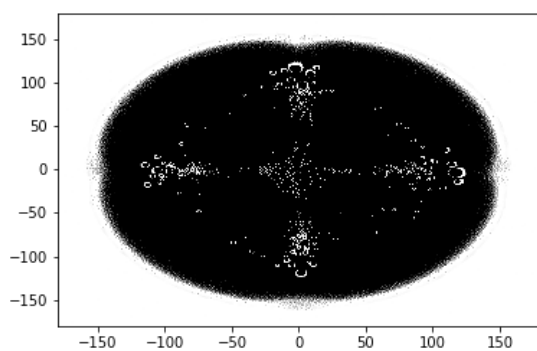**(32) Persian**

(33) Polish

(34) Punjabi

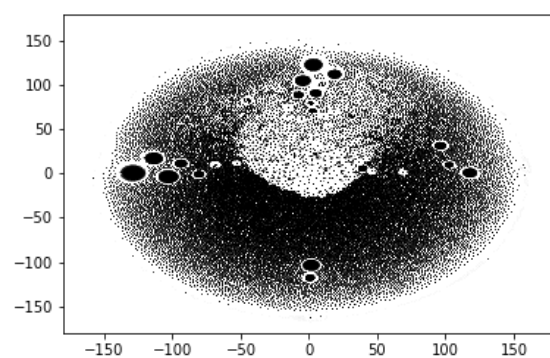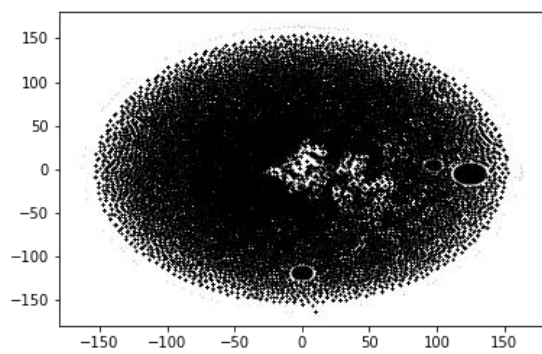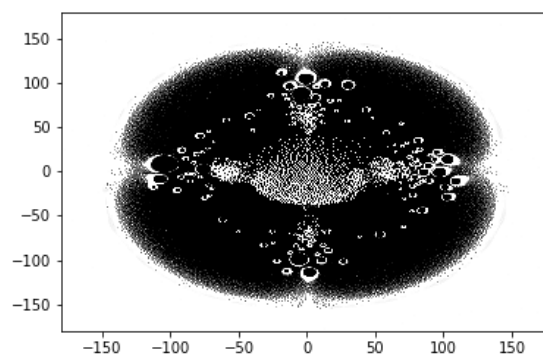**(35) Quechua**
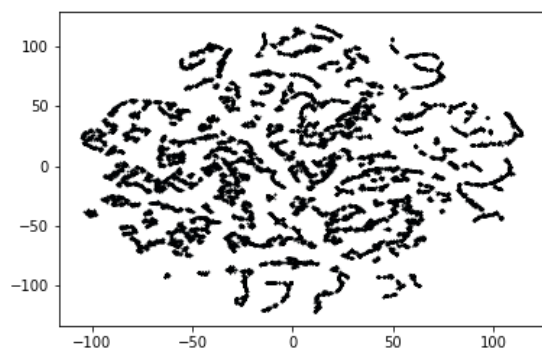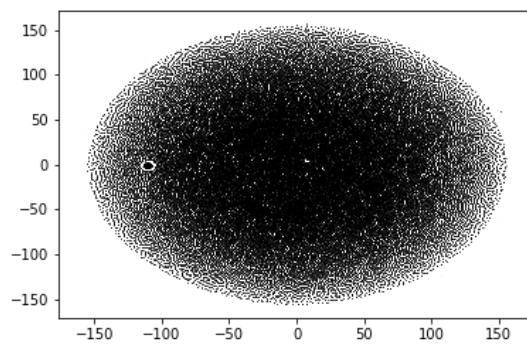

(36) Romanian

(37) Russian

(38) Serbian

(39) Sinhala

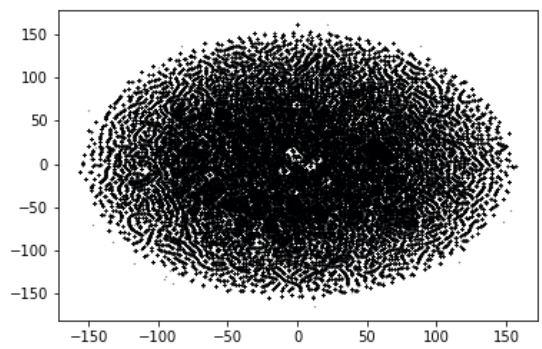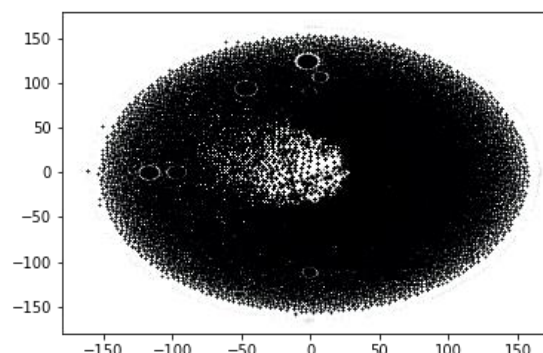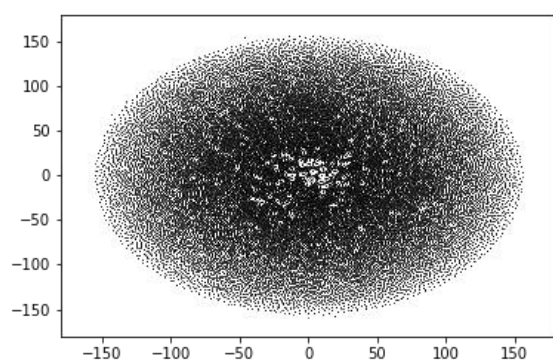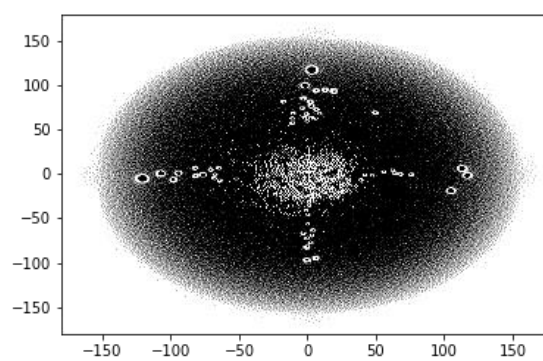**(40) Spanish**
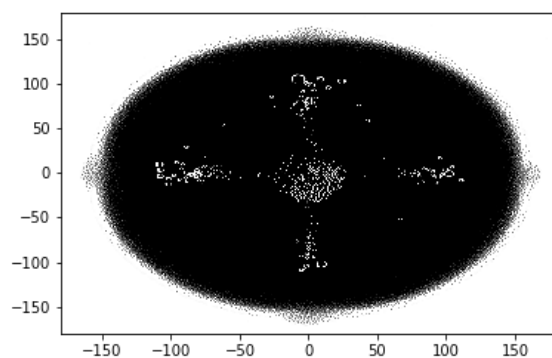

(41) Swahili
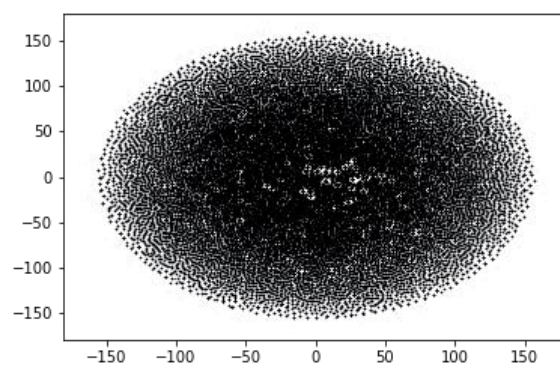

(42) Swedish

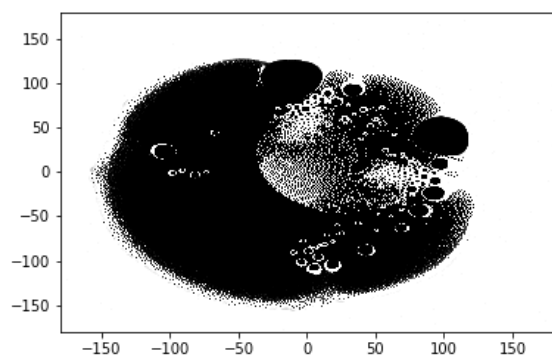**(43) Tabasaran**

(44) Tagalig
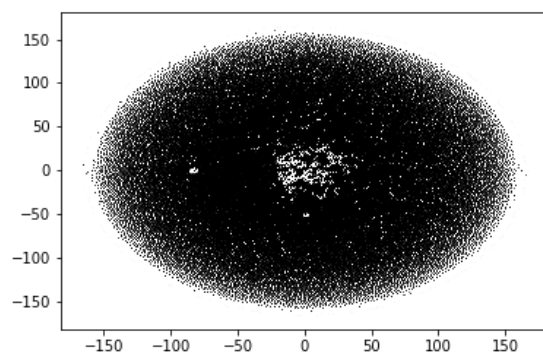

(45) Tatar

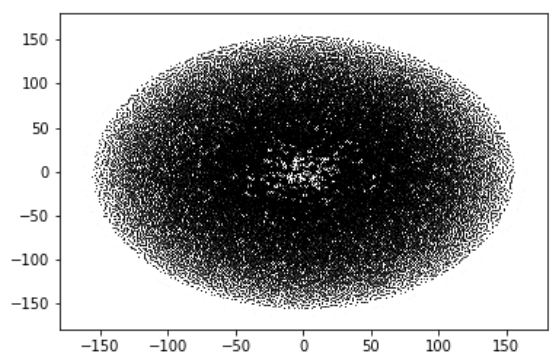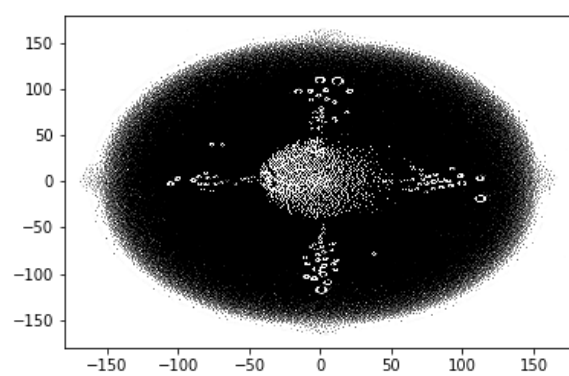**(46) Thai**

(47) Turkish

(48) Tuvan

(49) Udmurt

(50) Ukrainian

(51) Uzbek
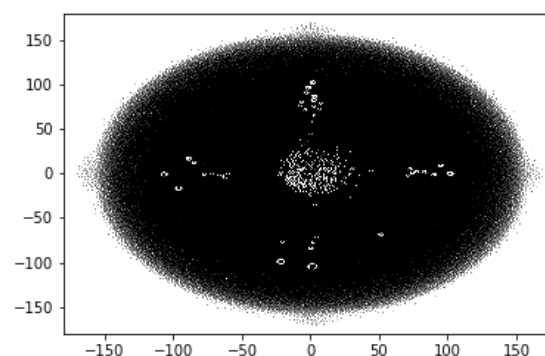
(52) Vietnamese

Fig. 9: t-SNE visualization of bigrams for 52 languages.