**TSNE**

**Unigram**

|  |  |
| --- | --- |
|  |  |
| 1. **Amharic** | 1. **Arabic** |
|  |  |
| 1. **Atikamekw** | 1. **Basque** |
|  |  |
| 1. **Belarusian** | 1. **Bengali** |
|  |  |
| 1. **Bulgarian** | 1. **Chechen** |
|  |  |
| 1. **Chinese** | 1. **Coptic** |
|  |  |
| 1. **Czech** | 1. **Dholuo** |
|  |  |
| 1. **Dutch** | 1. **English** |
|  |  |
| 1. **Esperanto** | 1. **Finnish** |
|  |  |
| 1. **French** | 1. **German** |
|  |  |
| 1. **Hindi** | 1. **Icelandic** |
|  |  |
| 1. **Indonesian** | 1. **Japanese** |
|  |  |
| 1. **Kabyle** | 1. **Kazakh** |
|  |  |
| 1. **Latin** | 1. **Hungarian** |
|  |  |
| 1. **Malayalam** | 1. **Navajo** |
|  |  |
| 1. **Norwegian** | 1. **Oromo** |
|  |  |
| 1. **Ossetian** | 1. **Persian** |
|  |  |
| 1. **Polish** | 1. **Punjabi** |
|  |  |
| 1. **Quechua** | 1. **Romanian** |
|  |  |
| 1. **Russian** | 1. **Serbian** |
|  |  |
| 1. **Sinhala** | 1. **Spanish** |
|  |  |
| 1. **Swahili** | 1. **Swedish** |
|  |  |
| 1. **Tabasaran** | 1. **Tagalig** |
|  |  |
| 1. **Tatar** | 1. **Thai** |
|  |  |
| 1. **Turkish** | 1. **Tuvan** |
|  |  |
| 1. **Udmurt** | 1. **Ukrainian** |
|  |  |
| 1. **Uzbek** | 1. **Vietnamese** |

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

**TSNE**

**Bigrams**

|  |  |
| --- | --- |
|  |  |
| 1. **Amharic** | 1. **Arabic** |
|  |  |
| 1. **Atikamekw** | 1. **Basque** |
|  |  |
| 1. **Belarusian** | 1. **Bengali** |
|  |  |
| 1. **Bulgarian** | 1. **Chechen** |
|  |  |
| 1. **Chinese** | 1. **Coptic** |
|  |  |
| 1. **Czech** | 1. **Dholuo** |
|  |  |
| 1. **Dutch** | 1. **English** |
|  |  |
| 1. **Esperanto** | 1. **Finnish** |
|  |  |
| 1. **French** | 1. **German** |
|  |  |
| 1. **Hindi** | 1. **Icelandic** |
|  |  |
| 1. **Indonesian** | 1. **Japanese** |
|  |  |
| 1. **Kabyle** | 1. **Kazakh** |
|  |  |
| 1. **Latin** | 1. **Hungarian** |
|  |  |
| 1. **Malayalam** | 1. **Navajo** |
|  |  |
| 1. **Norwegian** | 1. **Oromo** |
|  |  |
| 1. **Ossetian** | 1. **Persian** |
|  |  |
| 1. **Polish** | 1. **Punjabi** |
|  |  |
| 1. **Quechua** | 1. **Romanian** |
|  |  |
| 1. **Russian** | 1. **Serbian** |
|  |  |
| 1. **Sinhala** | 1. **Spanish** |
|  |  |
| 1. **Swahili** | 1. **Swedish** |
|  |  |
| 1. **Tabasaran** | 1. **Tagalig** |
|  |  |
| 1. **Tatar** | 1. **Thai** |
|  |  |
| 1. **Turkish** | 1. **Tuvan** |
|  |  |
| 1. **Udmurt** | 1. **Ukrainian** |
|  |  |
| 1. **Uzbek** | 1. **Vietnamese** |

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