Synergize the cross-lingual similarities to come up with a better language representation

- **Supervisor**: Ayyoob Imani
- **Examiner**: Prof. Schütze
- **BSc, MSc, Open**: MSc/BSc
- **Summary**: Since the success of deep learning, a number of models (like BERT and Word2vec) were presented that can represent natural language with a quality that is far better than any previously known models. A popular line of research after the introduction of these models was multilinguality. Here, the researchers tried to come up with models that are multilingual, i.e. models that can represent at least two languages at the same time. One way to do this is to take two unilingual models (like English word2vec and German word2vec) and try to align them together, which would yield a new model that can represent both German and English in the same space. Another way is to create models that are inherently multilingual, like MBERT, which represents all the languages in the same space. Many researchers have tried to use existing bilingual corpora (like parallel corpora) to create multilingual models or improve the quality of existing multilingual models. In this project, we will try to answer this question: “what if we have a multilingual corpus instead of a bilingual corpus?” while a bilingual corpus is a resource in two languages, a multilingual resource is in multiple languages. For example, if having a bilingual resource of English and German can help us to align German to English, would a resource of English, German, and Dutch help us to align German to English even better? In other words, can Dutch act as a bridge language to help English and German align even better?
- **Prerequisites**: enthusiasm, Good programming background (preferably python), basic knowledge of NLP, DL, and Pytorch