

Description de notre sujet d'ontologie

Ce projet de groupe porte sur les langues de la famille de langues indo-européennes, ces dernières couvrent principalement le continent européen, elles s'étendent cependant aux autres continents. L'ontologie décrira ainsi les familles de langues et langues issues de la famille de langues indo-européennes, décrivant ainsi les langues romanes, les langues slaves, les langues germaniques et les langues celtiques. Chacune de ces familles de langues comporte plusieurs langues et des spécificités qui peuvent permettre de les apparenter aux langues de leur famille, mais également de les différencier, par exemple, le ou les alphabets officiels peuvent permettre d'établir des relations entre les langues, comme le fait de posséder deux alphabets officiels (latin et cyrillique) pour le BCMS, une langue slave.

Utilisation des capacités d'OWL

OWL est un langage de description de connaissances, il permet de décrire des concepts, des propriétés et des relations entre ces concepts et propriétés. Il est donc possible de décrire des concepts et des propriétés dans un langage de description de connaissances et de les utiliser dans un langage de requête. C'est ce que nous allons voir dans cette section.

Raison du choix du sujet

Le choix de ce sujet s'est fait de part notre cursus universitaire qui tourne principalement autour des langues et de la linguistique. Il est intéressant de pouvoir allier la linguistique à l'informatique pour pouvoir manipuler différentes connaissances linguistiques à notre disposition et les modéliser. Un domaine où on se sentait à l'aise tous les 3, et qui nous permettait de faire un sujet assez complet. Les langues ont pas mal de propriétés, mais aussi de relations entre elles.

Points de départ

Pour déterminer les différentes propriétés, caractéristiques et éléments des différentes familles de langues choisies, nous nous sommes dans un premier temps basés sur nos expériences personnelles, nos intuitions, tout en prenant en exemple la langue qui nous était à tous familière, le français. Cela nous a permis de réaliser une première ébauche de description des concepts et les propriétés. Nous avons ensuite cherché des sources pour compléter notre description. Nous avons majoritairement utilisé des ressources en ligne, surtout pour les données chiffrées. Mais nous avons également utilisé des anciens supports de cours pour des relations comme adstrats, superstrats, et cætera.

Définition du domaine et de sa portée

L'idée est d'avoir une ontologie des langues indo-européennes.

L'objectif est de pouvoir décrire les relations entre les diverses langues, mais aussi de pouvoir décrire les propriétés des langues.

Les propriétés ont été définies en fonction de nos connaissances personnelles, dans le but de pouvoir également mettre en relation les langues en fonction de ces propriétés.

Par exemple:

- si deux langues ont une même langue mère, elles sont donc soeurs
- les langues ont un nombre de locuteurs, ainsi on peut interroger sur les langues les plus parlées, trier par nombre de locuteurs, etc.
- une langue peut être une langue officielle, ou une langue minoritaire, etc; on peut donc interroger par pays, par continent, etc.

Ainsi, un utilisateur pourrait chercher selon des propriétés ou des ensembles de propriétés, un peu à la manière d'une base de données, mais aussi se servir des inférences pour pouvoir générer des requêtes plus intuitives (par exemple, au lieu de lister les langues avec 0 locuteurs, on peut lister les langues mortes).

Ces mêmes inférences nous permettent également d'assurer la cohérence de notre ontologie.

Définition de la portée

L'ontologie est limitée aux langues indo-européennes, et aux propriétés et relations qui en découlent. La graphie, ainsi que les dialectes des langues ne seront pas prises en compte. Nous n'incluons seulement l'alphabet dont la langue se sert sans prendre en compte les caractères spécifiques à la langue comme les utilisations de dicritiques et de ligatures.

Dans quel contexte est utilisée l'ontologie

L'ontologie créée dans le cadre du projet du cours de Web Semantique sera utilisée pour décrire des concepts et des propriétés relatifs aux langues. Mais aussi et surtout pour interroger des données sur ces concepts et propriétés.

Le but est de pouvoir interroger des données sur les langues indo-européennes, et de pouvoir les comparer entre elles. Mais aussi de pouvoir requêter sur les propriétés des langues, comme le nombre de locuteurs, ou le nombre de pays où la langue est officielle pour obtenir la liste des objets répondant à ces critères.

On pourrait imaginer des requêtes comme:

- Quelles sont les langues qui ont un nombre de locuteurs supérieur à 100 millions?
- Quelles sont les langues qui ont moins de locuteurs que le français?
- Quelles sont les langues qui sont adtrats du français?
- Quelles sont les langues qui sont des langues vivantes?
- Quelles sont les langues qui sont utilisées dans plus de 5 pays?
- Combien de langues sont de type SVO, et combien sont de type SOV?
- Combien de locuteurs maternels a le français?
- Quelles langues sont ancêtres du français?
- Quelles sont les propriétés communes entre le français et l'allemand?