

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Interoperability and flexibility with language resources

Damir Ćavar

University of Zadar,
Institute of Croatian Language and Linguistics,
and Indiana University

Introduction

Language resources and computational linguistics technology over the last decades were usually developed from the perspective of specific research or application interests, with specific theoretical backgrounds and empirical or application oriented goals. While in the past the underlying technologies were often platform dependent and even hardware specific, with emerging trends in platform independent programming and software development environments (e.g. Python, Java, and even .NET), the technological aspects of language technology potentially could be, and often are, interoperable, i.e. platform and system independent. On the other hand, the language resources as such tend to be heterogeneous and highly specific.

With standards for data persistence, coding (e.g. Unicode-based), structuring and annotation, as for example provided by XML and related technologies, and specific annotation standards on the basis of such technologies (e.g. TEI and related standards), at the data level, interoperability seems to be moving a step forward towards better interoperability.

However, getting deeper into language resource annotation standards, e.g. TEI, one will potentially at some level face serious problems with the vast amount of tag and attribute choices that potentially even leads to variation within one resource where various annotators use the same standard and strategies. Cross-linguistic variation seems to hinder a uniform standard with a wide interoperability potential.

Under-specification and the freedom with respect to specific annotation possibilities causes a problem as well, where specific guidelines and best-practice examples are still missing. Linguistic annotation standards tend to be either too language specific, or they provided are very liberal and unspecific (e.g. morpho-syntactic annotation or part-of-speech tagging, or syntactic structure annotation).



The lack of standards for a more universal PoS-tagging, constituent structuring and annotation, the lack of mapping possibilities, of levels of granularity and maybe theoretically motivated annotation types to other schemata is a major issue for interoperability. Although various projects aim at providing more general or universal schemata for at least the morpho-syntactic level, and although there were attempts to develop a more theory neutral syntactic annotation, it seems that not much effort is put into specifications of annotation standards for meta levels for each linguistic level, and mappings between those. The consequences are that the annotation granularity of tagged corpora or tree banks is either geared towards specific theoretical and practical interests, or lacks the necessary granularity for interoperability.

There is definitely a need for a standards infrastructure that provides recommendations at this level, as well as mapping schemata, guidelines and samples.

