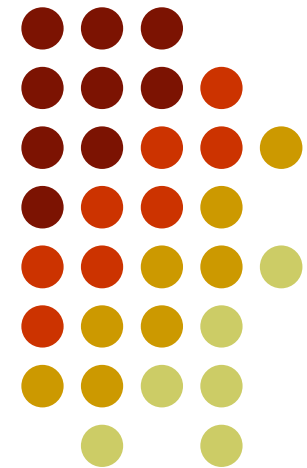


# Acquisition of domain-specific multiword expressions in Serbian

Vesna Pajić, Miloš Pajić

Center for data mining and bioinformatics  
University of Belgrade – Faculty of Agriculture



# From domain-specific MWEs to the terminology



- over 70% of the terms are composed of more than one word (Krieger and Finatto (2004)) – Multiword Terms (MWTs)
- domain-specific MWEs are difficult to detect automatically
- morphosyntactic analysis is used to improve the extraction of MWE
- for MWTs extraction, it is necessary to know the relevant syntactic structures used in particular language and the domain for expressing specific concepts

# Research specifics



## Objectives:

- Populate existing MWE dictionary of Serbian
- Starting the creation of agricultural terminology
- Analyse the properties of MWTs from agricultural domain in Serbian

## Tools and Resources:

- Collection of scientific papers in Serbian from the agricultural engineering domain
- Electronic morphological dictionaries for Serbian (simple words, multi-words)
- Unitex 3.0

# Results



716000 tokens in total

29,753 expressions of the A N structure (~12000 unique)

34% of the 1000 top frequent were MWTs from the agricultural engineering domain

***angažovana snaga*** (engaged power), ***aromatično bilje*** (aromatic plants), ***genetički potencijal*** (genetic potential), ***hidraulički sistem*** (hydraulic system)

***najveći broj*** (maximum number), ***nova tehnologija*** (new technology), ***dobijeni rezultat*** (obtained result)