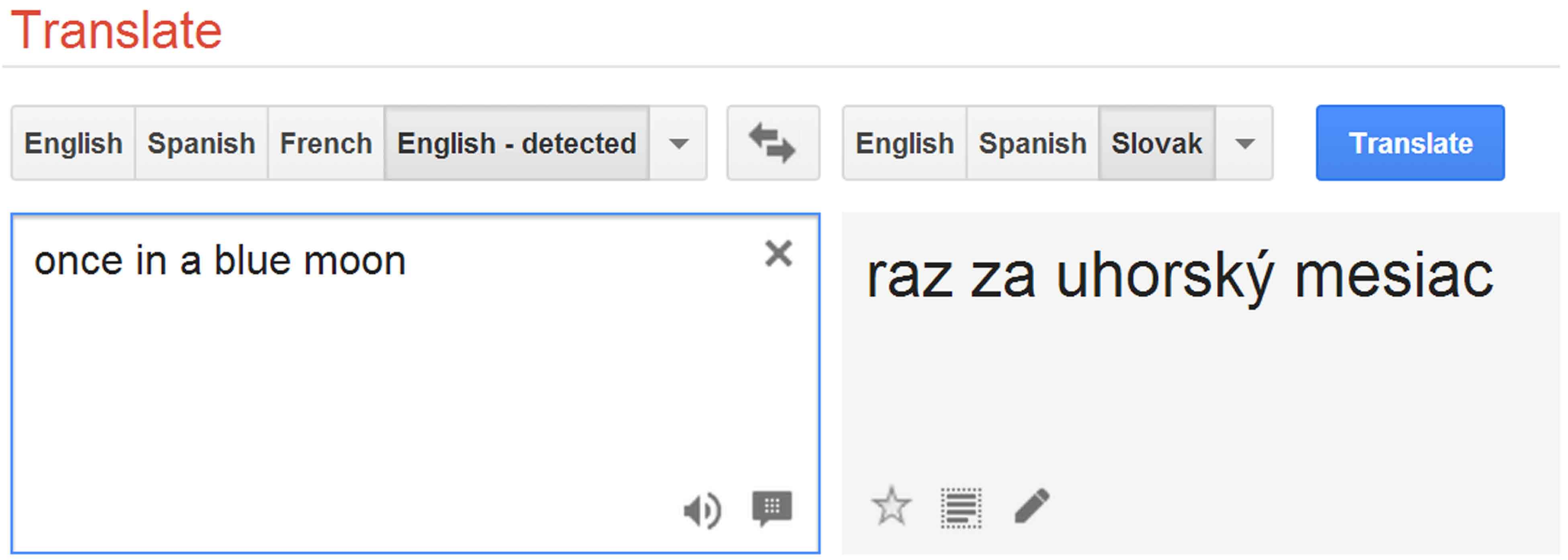
Collocation Extraction on the Web

Author: Martin Plank
Supervisor: Marián Šimko

Motivation

- identify groups of words with specific meaning
- improve machine translation, keyword extraction, natural language generation...



once in a blue moon

raz za uhorský mesiac

State-of-the-art

Approaches

1. **statistical** - association measures, e.g. pointwise mutual information:

$$PMIScore = \log \frac{F(xy)}{F(x) \cdot F(y)}$$

2. **linguistic** - based on collocation characteristic features:
   - limited compositionality, substituability and modifiability
   - not available for Slovak language
   - unsatisfactory results

Our approach

1. novel association measure
2. novel linguistic method

1. Novel association measure

- improved statistical measure PMI
- replace simple word frequencies with document frequencies (TF-IDF analogy):

$$DFScore = \log \frac{F(xy)}{DF(x) \cdot DF(y)}$$

2. Novel linguistic method

- based on **modifiability**

Headword supplements:
dlhý klinec, hrdzavý klinec...

Trafiť klinec po hlavičke

Candidate modifications:
trafiť dlhý/hrdzavý klinec po hlavičke...

- **observation:** in a collocation, candidate modifications have small frequencies, compared to the frequencies of headword supplements

- computing modifiability score:

$$S = \sum_{i=1}^{n} log \frac{F(M_i)}{F(Sup_i)}$$

- computing modifiability (using document frequencies):

$$ModDFScore = \frac{S \cdot \prod_{i=1}^{n} DF(w_i)}{NF \cdot F(c)}$$

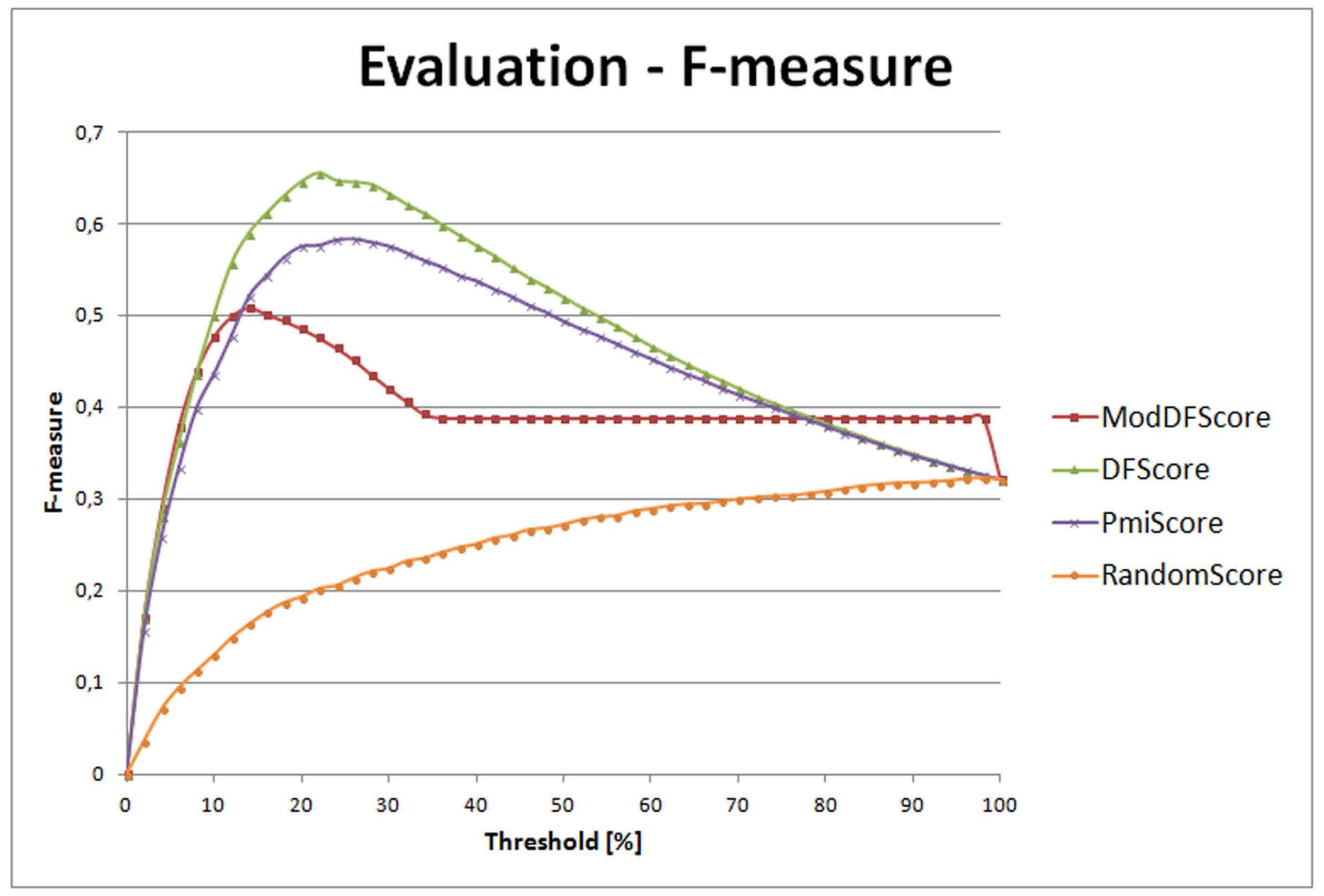
Conclusions

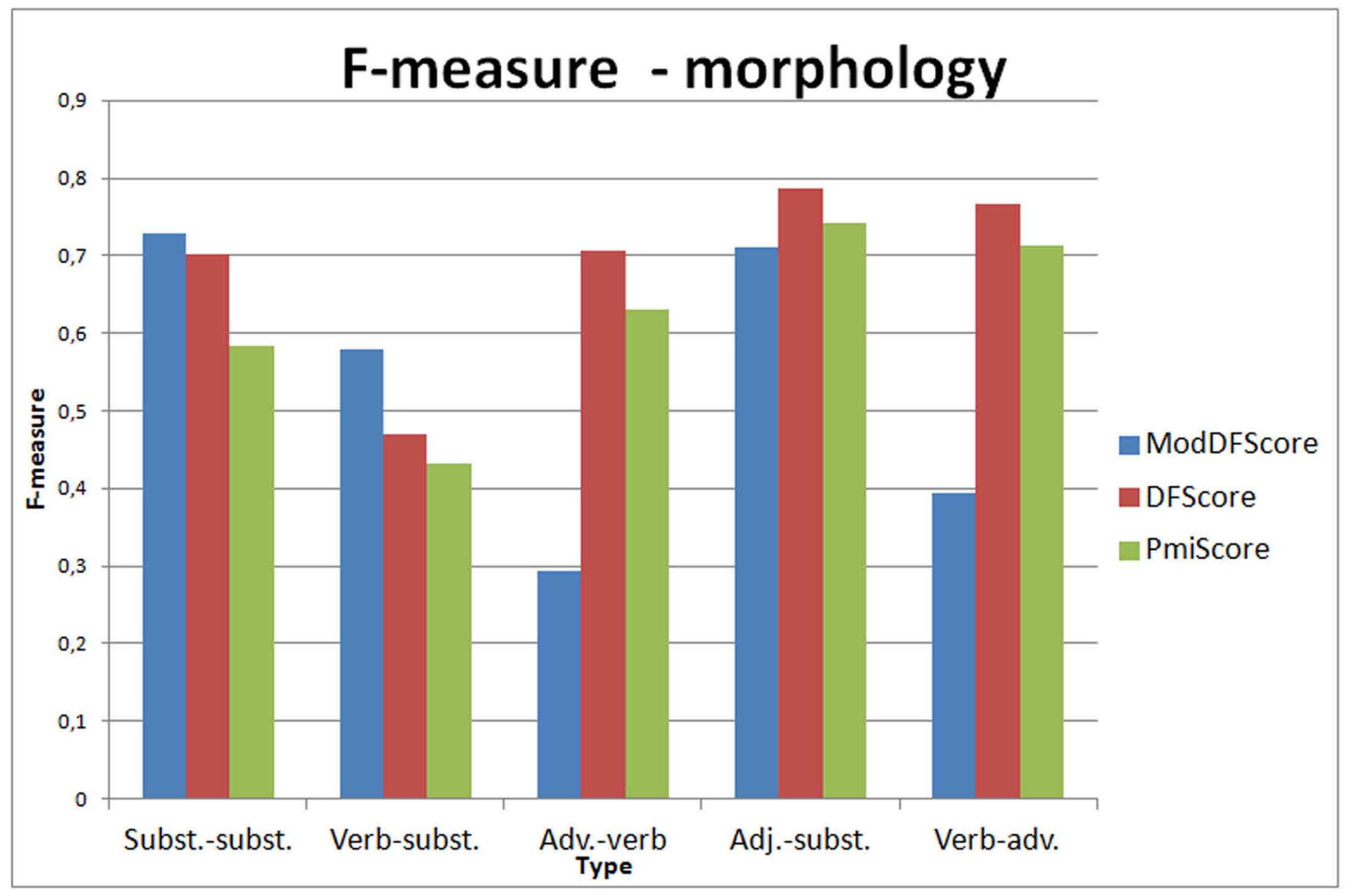
- **first study** of automatic collocation extraction in Slovak language
- proposal of **novel** statistical and linguistic method
- DFScore **out-performs** PMI (state-of-the-art)
- ModDFScore successfull mainly for **high-frequented** word combinations (potencial to perform better on larger corpora) and for some morphologic types
- ModDFScore improved (18 %) when considering only „modifiable" candidates, **out-performs** PMI for frequent (f>100) „modifiable" candidates
- web service for collocation extraction

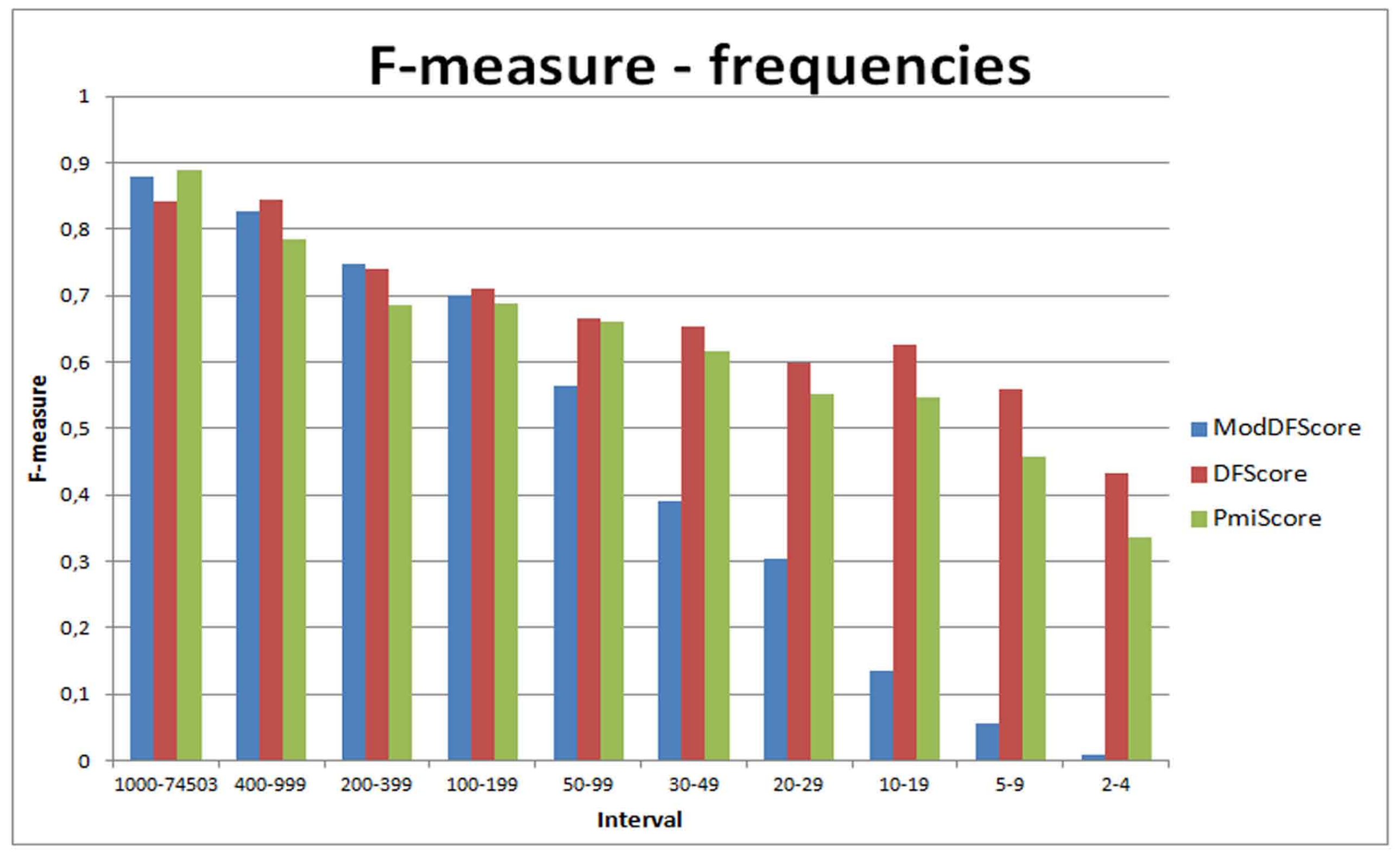
Evaluation

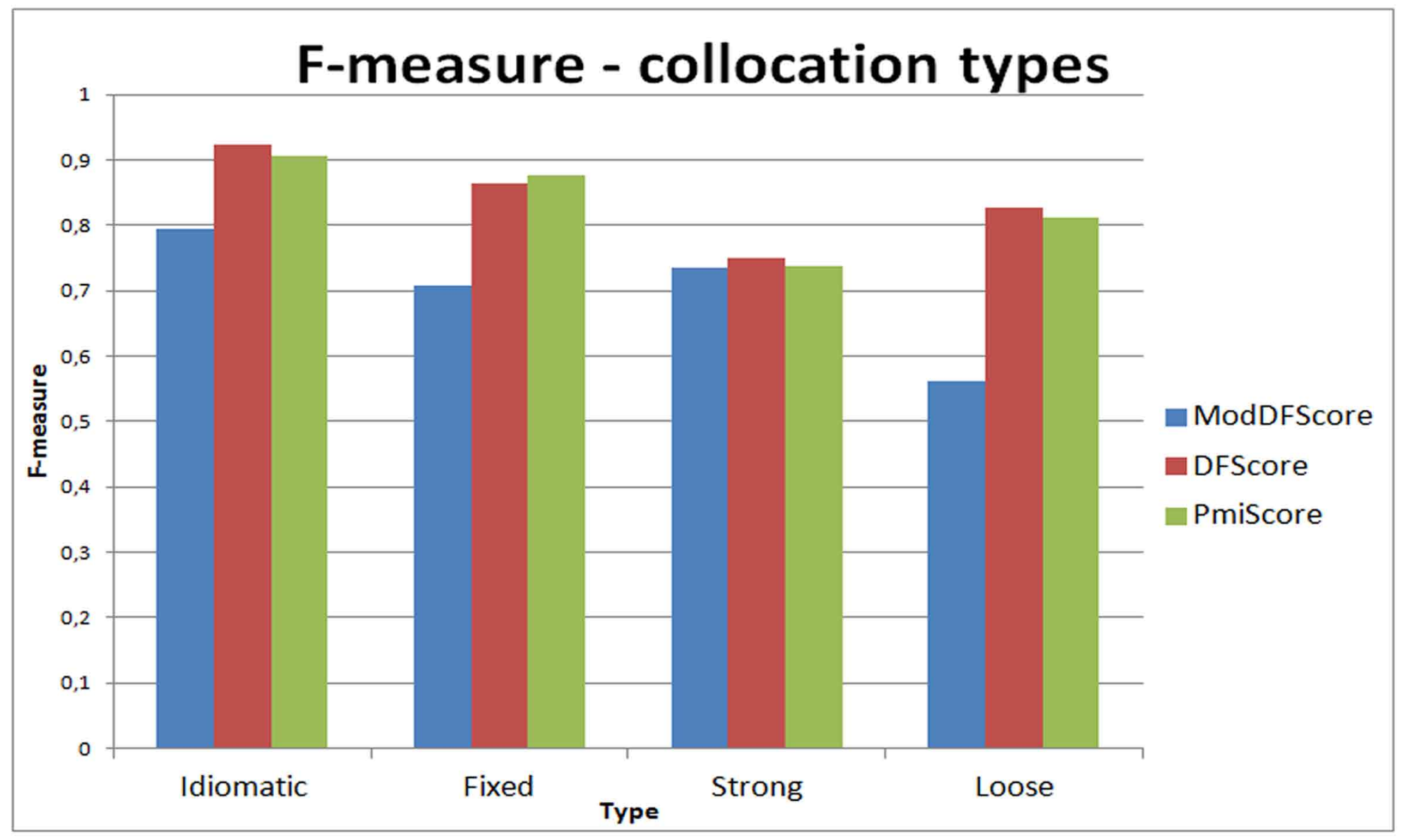
- detailed experiments
- dataset of slovak collocations (bigrams and trigrams)
- focus on frequencies, morphology and collocation types

Method	Precision	Recall	F-measure
PmiScore	0.51	0.69	0.58
DFScore	**0.61**	**0.71**	**0.66**
ModDFScore	0.60 (0.60)	0.44 (0.80)	0.51 (0.69)


Evaluation - F-measure


F-measure - morphology


F-measure - frequencies


F-measure - collocation types