



Extension of SNOMED CT resources for useful adoption in Slovak language

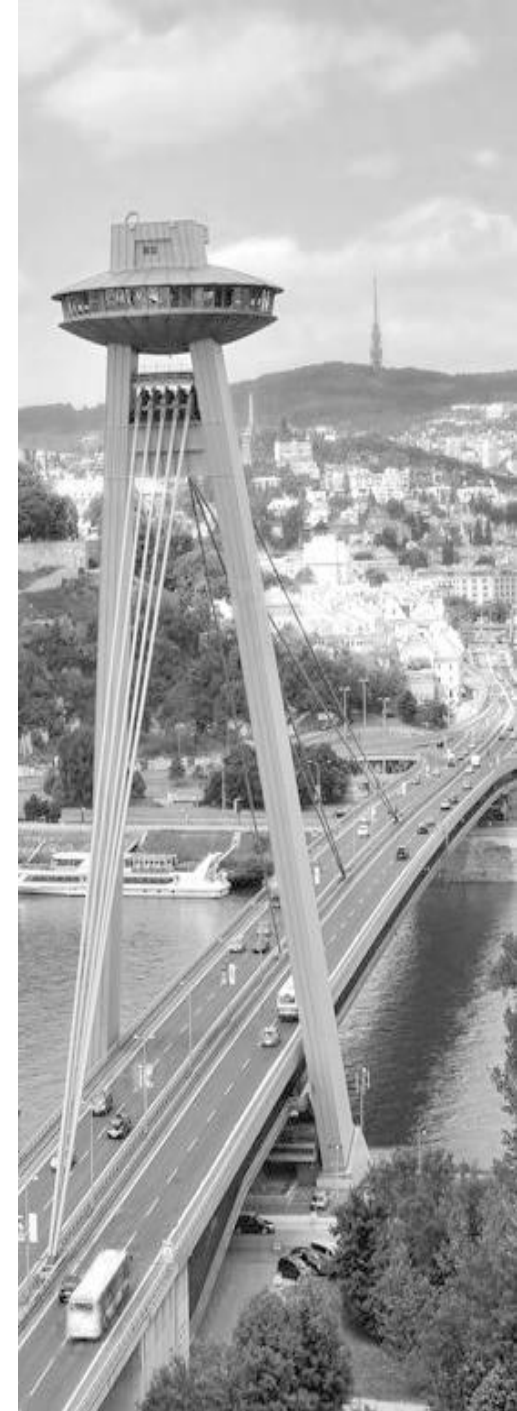
Milada Kovárová
PosAm spol. s r.o.

milada.kovarova@posam.sk

Jana Levická
Ľ. Štúr Institute of
Linguistics, SAS

janal@korpus.juls.savba.sk

PosAm



Clinicians use native language

Slovak language

Pacientka trpí zimnicou a hnačkami. Matka udáva zvýšenú teplotu do 38,5° trvajúcu dva dni.

Prejavuje sa slabý suchý kašeľ s mierne sťaženým dýchaním. Na vlasovej časti sú pozorované vyrážky. Podozrenie na infekciu ovčímí kiahňami. Odporúčaná liečba antipyretiká, Fenistil kvapky, kludový režim. (Očkovanie proti ovčím kiahňam MMRV 08/2014.)

Why is Slovak language complex?



Inflectional language

Modification of a word to express different grammatical categories (tense, case, person, number, gender, mood).



Complex morphology

6 cases, 3 genders and 2 numbers of nouns, 3 persons, 2 numbers of verbs, various tenses.



Variable word order

Strong inflection enables the identification of grammatical roles (subject, object, predicate, etc.) regardless of word placement.



Double negative

Two forms of negation used in the same clause are correct form of expressing negation.



Omitted subjects

If subject personal pronouns are not emphatic they are omitted.



Stem mutation

Changes within a word that indicate grammatical information (e.g. irregular verbs).

Pacientka trpí zimnicou a hnačkami. Matka udáva zvýšenú teplotu do 38,5° trvajúcu dva dni.

Prejavuje sa slabým sťaženým dýchaním, pozorované vyrážky, ovčímí kiahňami. antipyretiká, Fenistil kvapky, kludový režim. (Očkovanie proti ovčím kiahňam MMRV 08/2014.)

„zimnicou“ -

N: zimnica

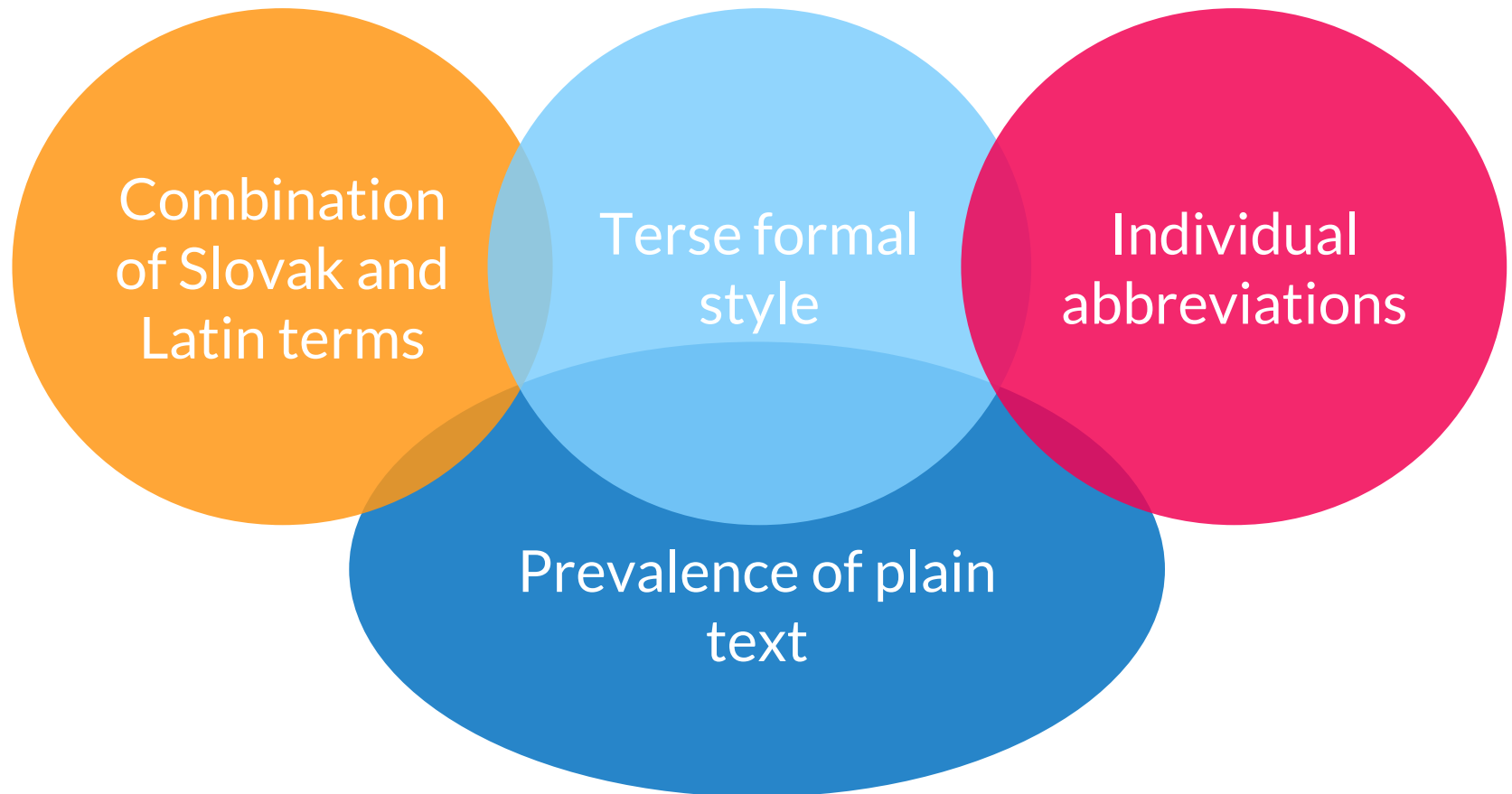
en. chill

Pacientka trpí zimnicami a
udáva zvýšenú teplotu
dvi.

Prejavuje sa slabý suchý kašeľ s mierne
sťaženým dýchaním. Na vlasovej časti sú
pozorované vyrážky. Podozrenie na infekciu
ovčimi kiahňami. Odporúčaná liečba
antipyretiká, Fenistil kvapky, kludový režim.
(Očkovanie proti ovčím kiahňam MMRV
08/2014.)

**„ovčimi kiahňami“ -
N: ovčie kiahne
en. varicella**

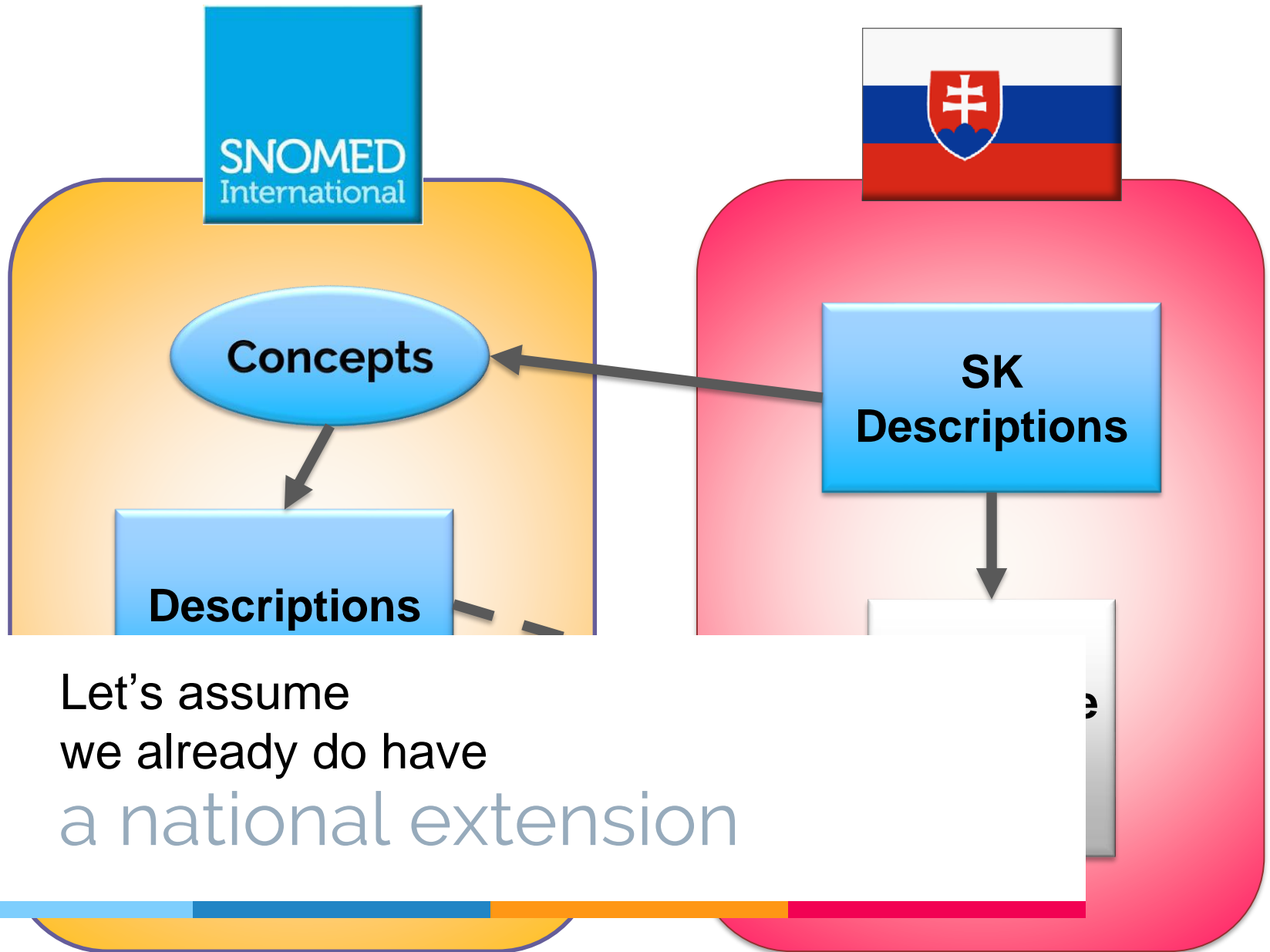
Plus medical language

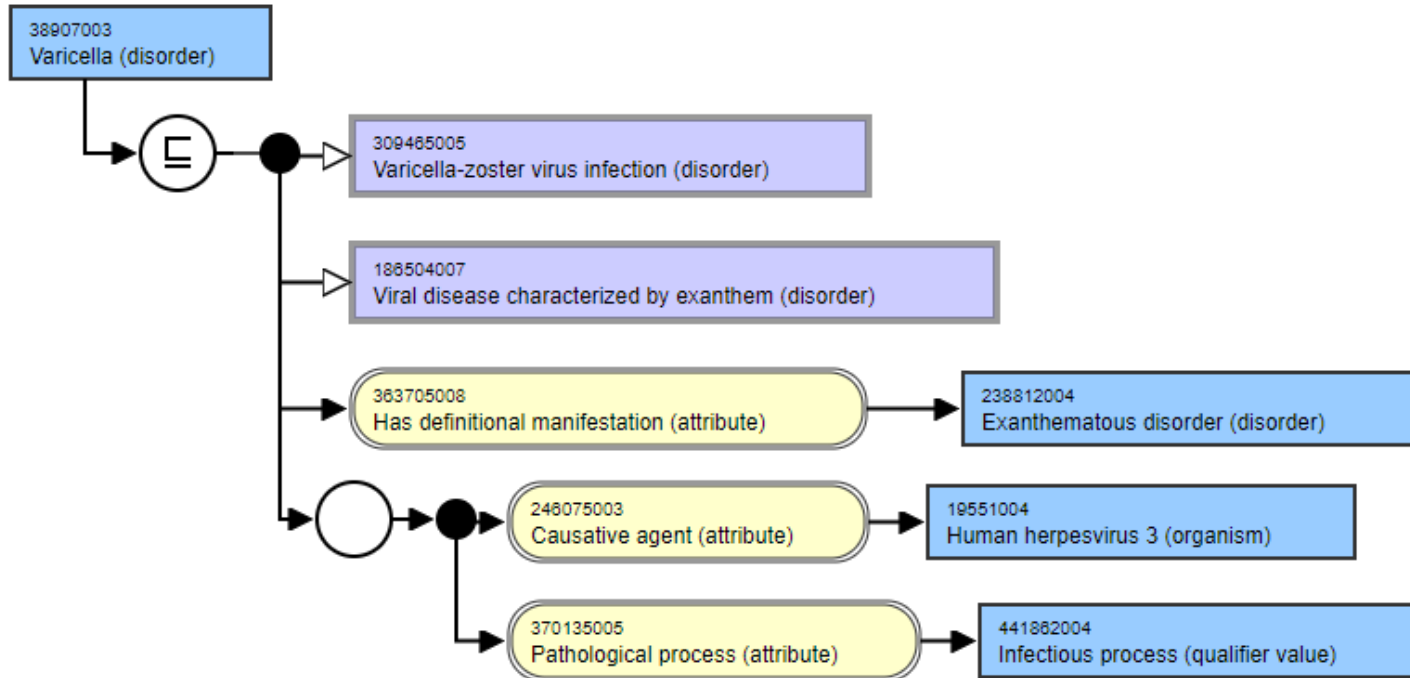




Big question:

***What can the SNOMED CT
offer us?***





	Term	Acceptability (US)
F	Varicella (disorder)	Preferred
S	Varicella	Preferred
S	Chicken pox	Acceptable
S	Varicella infection	Acceptable



Can we use it?

Basic target:

**search for SNOMED CT terms in clinical
documentation**

Pacientka trpí zimnicou a hnačkami. Matka udáva zvýšenú teplotu do 38,5° trvajúcu dva dni.

Prejavuje sa slabý suchý kašeľ s mierne sťaženým dýchaním. Vo vlasovej časti sú pozorované vyrážky. Podozrenie na infekciu ovčímí kiahňami. Odporúčaná liečba antipyretiká, Fenistil kvapky, kl'udový režim. (Očkovanie proti ovčím kiahňam MMRV 08/2014.)

Linguistic world offers

Resources and tools to be re-used

Morphological database of the Slovak language



Lemma „ucho“ (ear)

- ▷ Contains lemmas and their inflected word forms
- ▷ Used for automatic annotation of the Slovak National Corpus

 99.000 lemmas with full paradigm template

 1.100.000 unique word forms

 4.000.000 tokens of medical resources

SSns1: **ucho**

SSns2: **ucha**

SSns3: **uchu**

SSns4: **ucho**

SSns5: **ucho**

SSns6: **uchu**

SSns7: **uchom**

SSnp1: **uši**

SSnp2: **ušú uší**

SSnp3: **ušiam**

SSnp4: **uši**

SSnp5: **uši**

SSnp6: **ušiach**

SSnp7: **ušami**

Text pre-processing

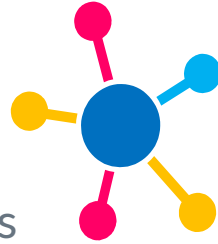
● → Segmentation

● → Tokenization

● → Stemming and
lemmatization

● → Language
modelling

Text processing



Rule-based

- ▷ Strong linguistic analysis
- ▷ Formal grammars
- ▷ Hand-made dictionaries

- ▷ Need for quality input data
- ▷ Guaranteed outputs

- ▷ Lower performance

- ▷ Suitable for closed domains

Statistical

- ▷ Text decomposition
- ▷ Token annotation
- ▷ Trained on real data

- ▷ Resistance to unexpected inputs
- ▷ Restricted guarantee

- ▷ Higher performance

- ▷ Generic approach



Text processing

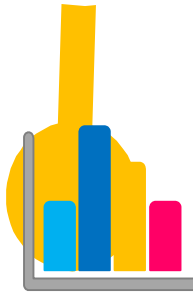
Rule-based

- ▷ Strong linguistic analysis
- ▷ Formal grammars
- ▷ Hand-made dictionaries
- ▷ Need for quality input data
- ▷ Guaranteed outputs
- ▷ Lower performance
- ▷ Suitable for closed domains

Statistical

- ▷ Text decomposition
- ▷ Token annotation
- ▷ Trained on real data
- ▷ Resistance to unexpected inputs
- ▷ Restricted guarantee
- ▷ Higher performance
- ▷ Generic approach

Combination



Extension of the extension

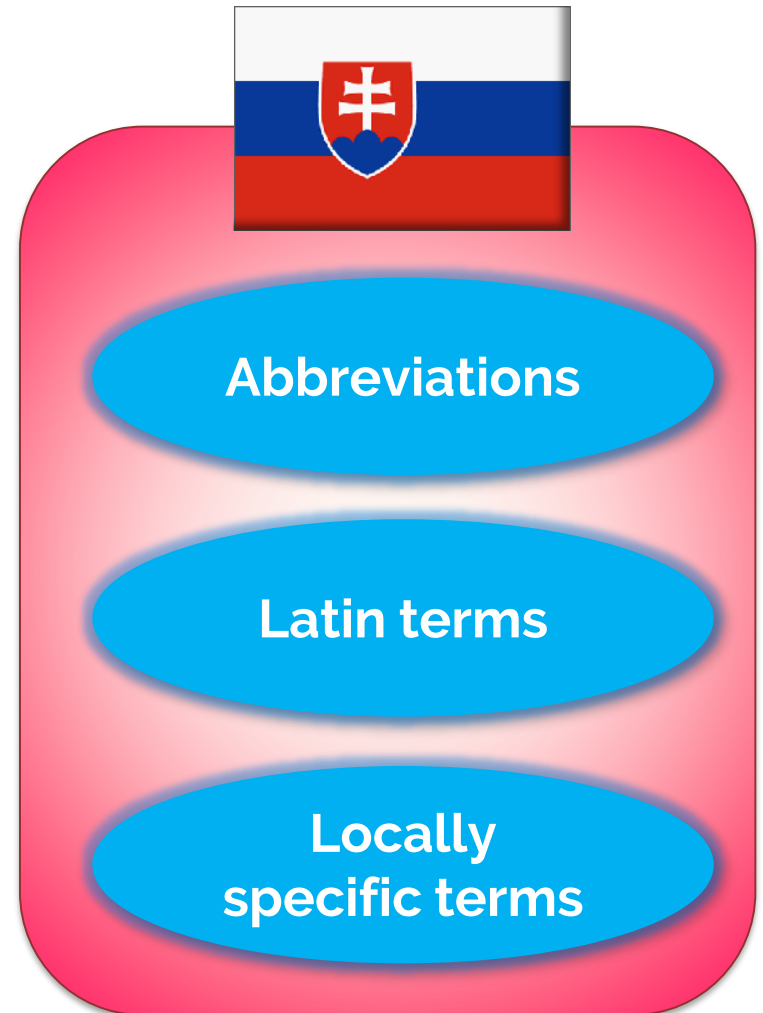
Resources and tools needed

Set of specific terms

Not contained in standard
SNOMED CT components

BUT

frequently used in medical
texts



SNOMED CT as a „dictionary“

1 Language extension

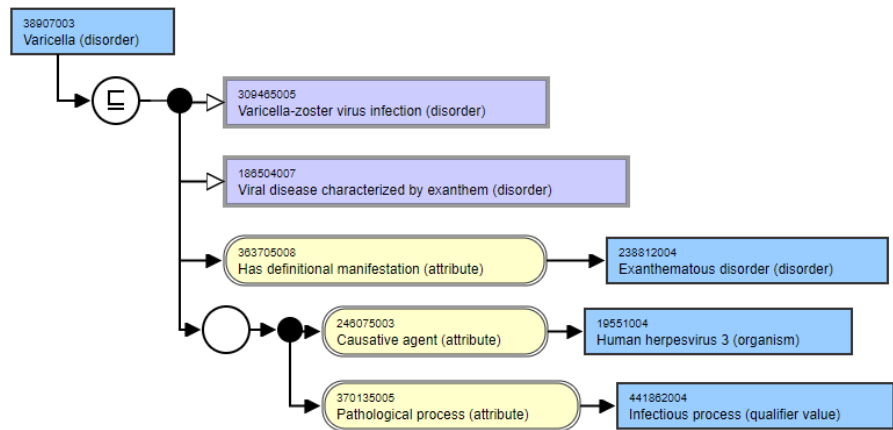
- ▷ Multilingual dictionary
- ▷ Requires grammar knowledge

2 Pairing with morphological database

- ▷ All possible forms of a term
- ▷ Requires rules for multi-word terms

3 Corpora analysis

- ▷ Adds usage frequency
- ▷ Requires additional context information



Term		Acceptability (SK)
F	ovčie kiahne (ochorenie)	Preferred
S	ovčie kiahne	Preferred
S	varicella	Acceptable



ns1	ovčia kiahňa	0,035
ns2	ovčej kiahne	0,001
...
np1	ovčie kiahne	0,672
np2	ovčích kiahní	0,102
np3	ovčích kiahňach	0,145
...

Rules

1 Descriptions denoted by multiple words

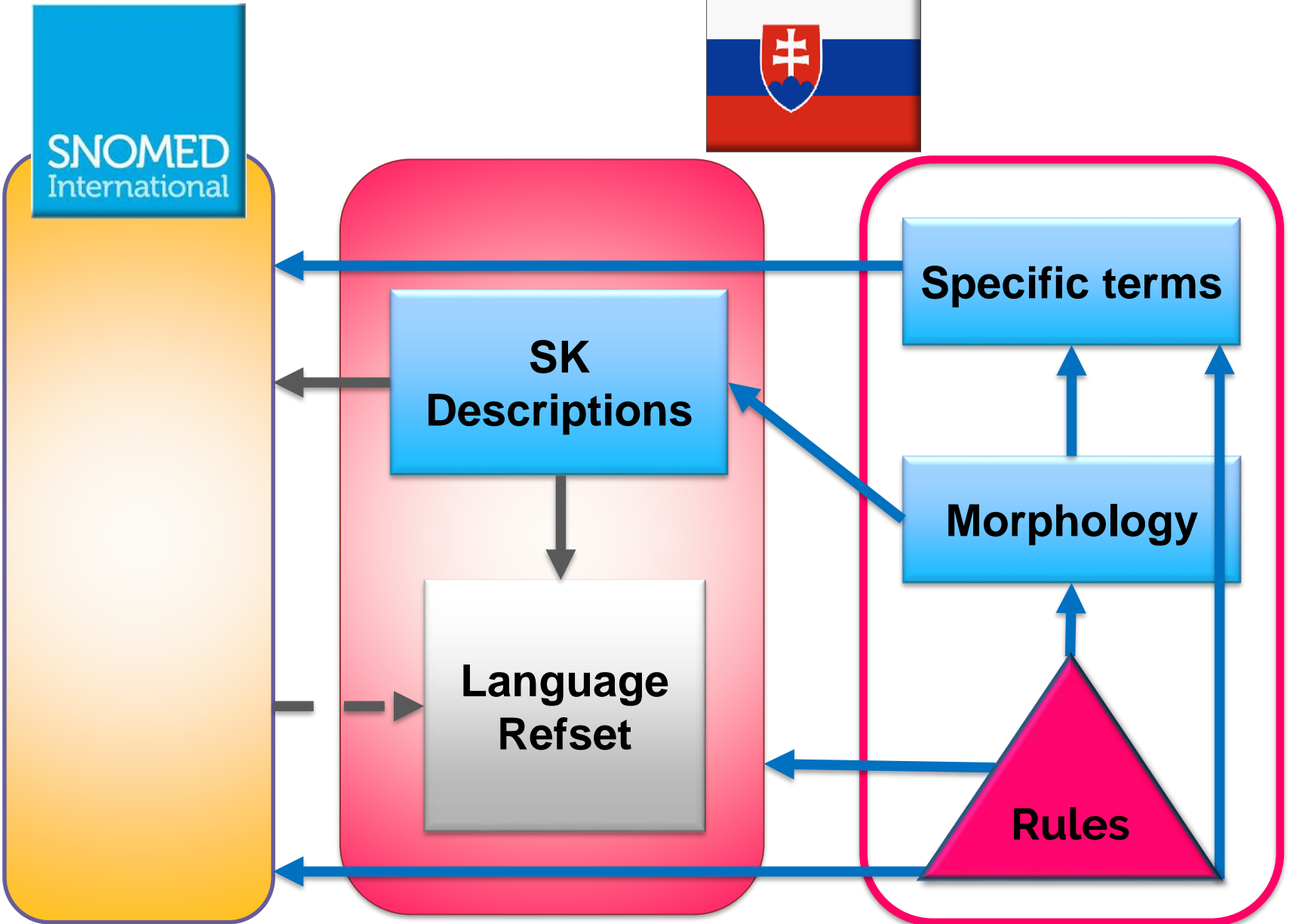
- ▷ Common occurrence
- ▷ Patterns

2 Semantic relationships (ontology)

- ▷ Causality
- ▷ Domain and range of attributes

3 Context

- ▷ Document meta-data
- ▷ Semantic tags of concepts or explicit context



Just the first step

Ambitions

Possibilities of further use

 **Automatic document annotation**

 **Extraction of structured data**

 **Machine translation**

 **Speech recognition**

Thanks!

Any questions?

You can find us at:

milada.kovarova@posam.sk

janal@korpus.juls.savba.sk