Material for the Semantic Collocations task

Task	Decide whether the given word pair forms a semantic collocation
Objects	Word pairs (identified by two lemmas L1 and L2)
Target class	YES / NO

Basic statistics (preprocessed data)

attribute	type	definition
L1	**string**	lemma of the first word
L2	**string**	lemma of the second word
A	**integer**	frequency of the bigram (L1, L2)
B	**integer**	frequency of all bigrams of the type (L1, ¬L2)
C	**integer**	frequency of all bigrams of the type (¬L1, L2)
D	**integer**	frequency of all bigrams of the type (¬L1, ¬L2)

Feature vector

feature	type	definition
A1	**continuous**	Pointwise mutual information
A2		First Kulczynski coefficient
A3		Unigram subtuples measure
A4		Ngram word coocurrence
A5		Reverse confusion probability
A6		Reverse cross entropy
A7		Right context phrasal entropy
A8		Log frequency biased mutual dependency
A9		Cosine context similarity in boolean vector space
A10		Dice context similarity in tf.idf vector space
A11	**categorical**	POS1:POS2

Legend

L1	lemma of the first word
L2	lemma of the second word
¬L1	a lemma different from L1
¬L2	a lemma different from L2
POS1	part of speech of L1
POS2	part of speech of L2

Barbora Hladká, Martin Holub ESSLLI 2013

What are semantic collocations

Material for the Semantic Collocation task

If one word collocates with another, they often occur together. Most generally, the term **collocation** denotes a meaningful word combination that often (regularly or frequently or typically) occurs in natural language.

Semantic collocations, in addition, form semantic units. Semantic collocations are multiword expressions that are lexically, syntactically, pragmatically and/or statistically **idiosyncratic**. It means, that semantic collocations have semantic and/or syntactic properties that cannot be fully predicted from those of their components, and therefore semantic collocations **have to be listed in a dictionary**.

- **Examples of semantic collocations**

example	translation	description
Masarykův okruh	Masaryk circuit	motor sport race track named after the first president of Czechoslovakia, Tomáš Garrigue Masaryk
trestní čin	criminal act	harmful act
šedá ekonomika	gray market	legal trade of a commodity through unofficial distribution channels
Antonín Dvořák	Antonin Dvorak	Czech composer
skleníkový efekt	greenhouse effect	it keeps the Earth's climate warm and habitable
trest smrti	death penalty	person is put to death by the state as a punishment for a crime.
znaková řeč	sign language	language which, instead of sound patterns, uses body language
zelená karta	green card	ID card attesting to the permanent resident status of an immigrant in the United States
rovnoramenný trojúhelník	isosceles triangle	triangle with two sides equal in length
řidičský průkaz	driving license	document stating that a given person can operate a motorized vehicles

- **Examples of simple collocations**

  high mountains, tall trees, strong tea, powerful car, heavy smoker, big mistake, light wind, the rich and famous, they are, do exercise.