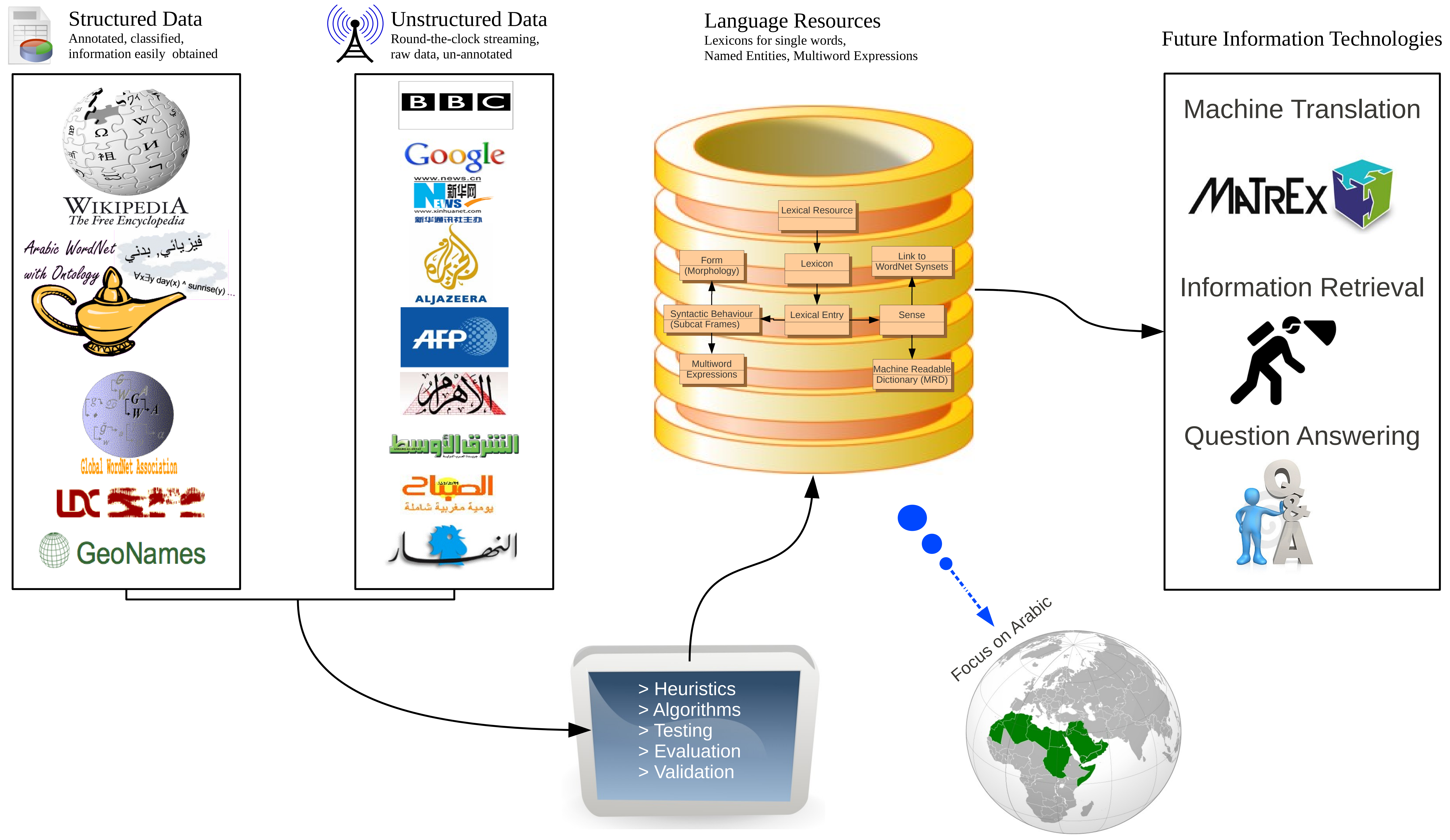
Construction of Language Resources for Enhancing Future Information Technologies

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, Josef van Genabith
School of Computing, Dublin City University



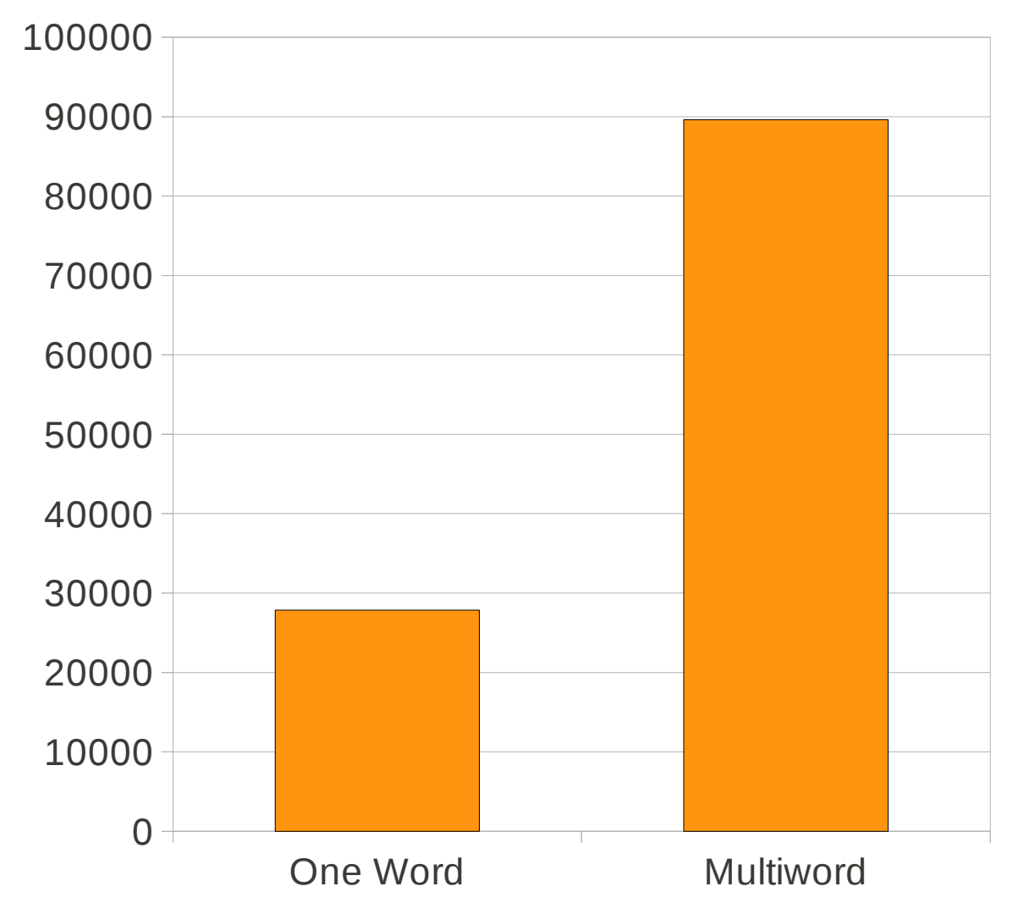
What are Multiword Expressions?

Multiword expressions are phrases that are composed of two or more words, and the meaning of the expression is not derived directly from the individual words.

Examples:
New York, black hole, life guard, touch screen, mobile phone, Windows Vista, Screen saver, etc.

Importance of Multiword Expressions

1. High frequency in natural text
   Search logs, technical entries, Wikipedia entries are mostly multiword expressions
2. Diversity. They have different forms
   Named Entities: *Saudi Arabia, Red Sea*
   Foreign forms: *ad hoc, carte blanche*
   Static: *by and large, easy money*
   Inflected: *jump/jumps/jumped to conclusion/conclusions*
3. Low ambiguity
   life => 13 meanings (Longman Dictionary of Contemporary English)
   guard => 5 as noun + 4 as verb
   life guard => 1 meaning
4. Statistically significant co-occurrence
   general elections, 2.4 million hits in Google
   public elections, 120,000 hits



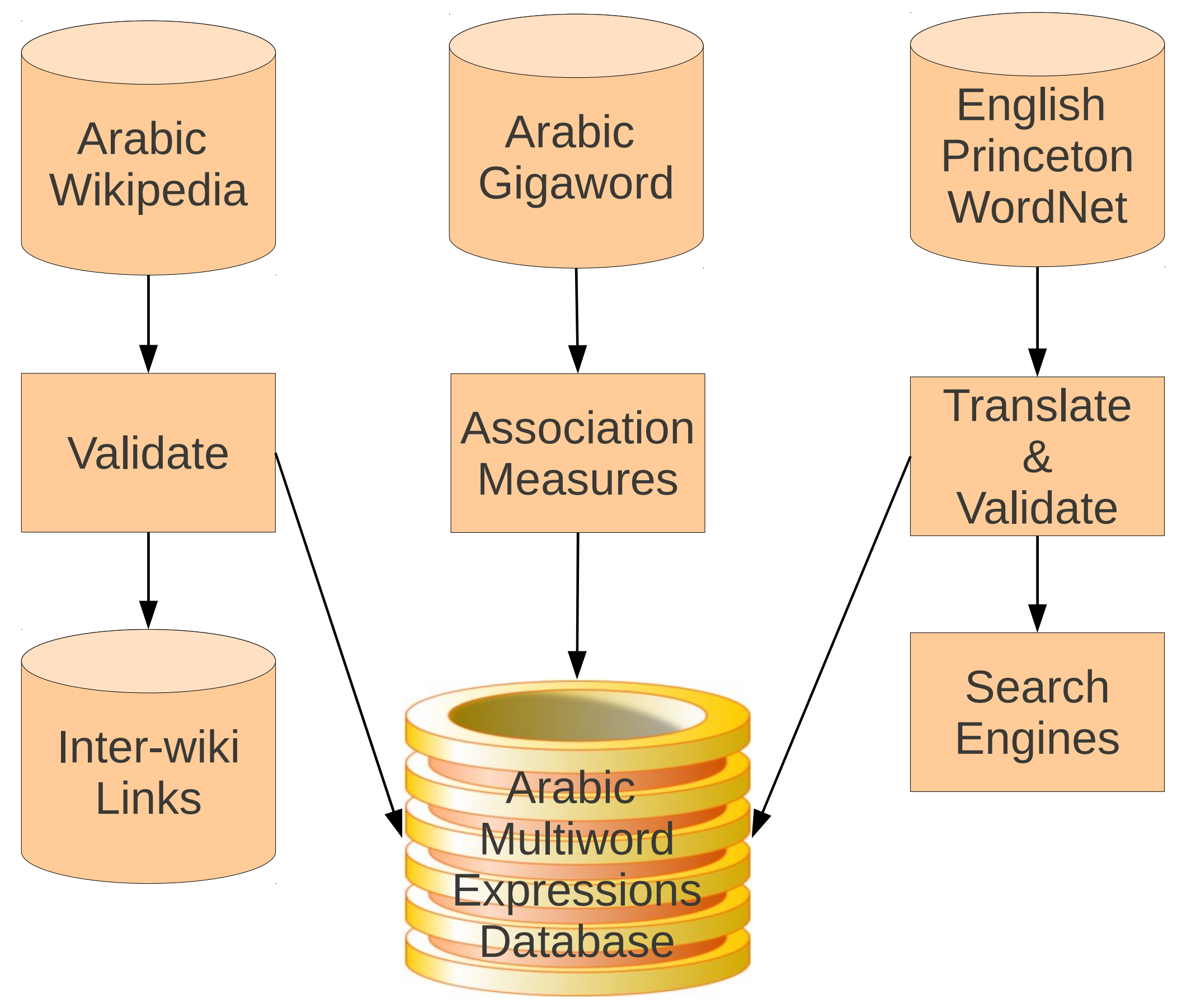
Frequency Ration of one word titles to multiword titles in the Arabic Wikipedia

Types of Multiword Expressions

1. Idiomatic
   hot dog, pelican crossing, blind alley, lame duck, ad hoc, vicious circle
2. Semi-Idiomatic
   fire engine, car boot sale, estate agent, easy money, black hole
3. Non-Idiomatic
   police station, traffic light, dish washer, general elections, washing machine

Non-multiword Expressions
office supplies, international affairs, planning centre, morning exercise

Pipeline of the Automatic Extraction of Multiword Expressions



Language ranking according to compounding

Language	Many-to-One	Language Family
German	65.01%	West Germanic
Swedish	64.51%	North Germanic
Danish	63.40%	North Germanic
Norwegian	62.77%	North Germanic
Dutch	60.41%	West Germanic
Russian	53.69%	East Slavic
Esperanto	50.13%	Artificial
Greek	34.98%	Indo-European
Czech	34.33%	West Slavic
Portuguese	33.67%	Romance
Turkish	33.33%	Turkic
Catalan	33.00%	Romance
French	32.95%	Romance
Polish	31.20%	West Slavic
Latin	30.71%	Italic
Italian	30.66%	Romance
Hebrew	29.98%	Semitic
Romanian	28.98%	Romance
Spanish	28.46%	Romance
English	28.11%	West Germanic
Indonesian	26.76%	Austronesian

Examples of Multiword Expressions

Arabic Phrase	Translation	# of languages	Ratio of M-2-1
فقر دم	anemia	21	100%
التهاب القولون	colitis	12	92%
ورق الحائط	wallpaper	11	82%
قمرة القيادة	cockpit	17	76%
فريق عمل	teamwork	9	67%
فرس النهر	hippopotamus	21	52%
قاعدة بيانات	database	20	45%
فرشاة أسنان	toothbrush	19	37%
بطاقة ائتمان	credit card	17	35%
إسعافات أولية	First aid	17	24%
فوهة بركانية	volcanic crater	14	21%
فن تجريدي	abstract art	20	15%
دائرة كهربائية	electrical network	20	5%
تاريخ الطيران	aviation history	12	0%