Facts and Figures: Languages in the Corpus

The following languages and varieties were annotated in the SMS. Please check our methodology for annotating languages to fully understand these figures.

Variety/language	Abbreviation	
German		
Standard German	deu	7'287
Swiss German	gsw	10'706
Other German	gda	9
French		
Standard French	fra	4'619
French Patois	fsw	30
Italian		
Standard Italian	ita	1'471
Italian Dialect	isw	48
Romansh		
Sursilvan	roh-sr	425
Sutsilvan	roh-st	9
Surmiran	roh-sm	110
Puter	roh-pt	181
Vallader	roh-vl	337
Grischun	roh-gr	59
Other languages		
English	eng	535
Dutch	nld	5
North Germanic	gmn	3
Slavic	sla	42
Spanish	spa	43
Portuguese	por	5
Modern Greek	gre	3
Arabic	ara	1
Other	oth	106

Please keep in mind that one SMS can have more than one main language, so if you add those figures together, you will get more than 100%. As you can see, some languages were summarized. If we say that an SMS was written in North Germanic, it can be Danish, Norwegian or Swedish. Because the individual SMS are so short, they often contain words that are pronounced in a similar way in more than one of those languages and because of the unorthodox spelling in the SMS we cannot rely on spelling either when defining languages. We thus decided to pull these languages together. The same goes for Slavic languages.