

MultilingualWeb Workshop Riga, 29 April 2015

Semi-automatic generation of multilingual glossaries



KDICTIONARIES

Ilan Kernerman
K Dictionaries Ltd, Tel Aviv

SUMMARY

K Dictionaries' semi-automated multilingual glossaries stem from our unique English multilingual dictionary:

- (1) reverse engineer parts of the initial data
- (2) edit the word lists and links and re-process the results
ready for 15 languages: with 43 languages each
- (3) expand with Linked Data & Semantic Web technologies
kicking off: lemon-based

The glossaries serve to deal with multilingual contents on the Web and to interconnect dozens of languages.

K DICTIONARIES *TechnologyDrivenContent*

- ▶ Multi-language/multi-layer content for 50 languages
 - *monolingual, bilingual & multilingual datasets*
 - *resources for language learning & translation*
 - *morphology & pronunciation, tools & applications*
- ▶ Established in 1993, based in Tel Aviv
- ▶ Cooperation with technology, publishing & academic partners worldwide

LINGUISTIC

- ▶ macro & microstructure
- ▶ editorial & translation styleguides
- ▶ metalanguage conversion tables
- ▶ headword & word form lists
- ▶ content & format revisions
- ▶ L1 lexicographer teams & L2 translators
- ▶ technical infrastructure synchronization

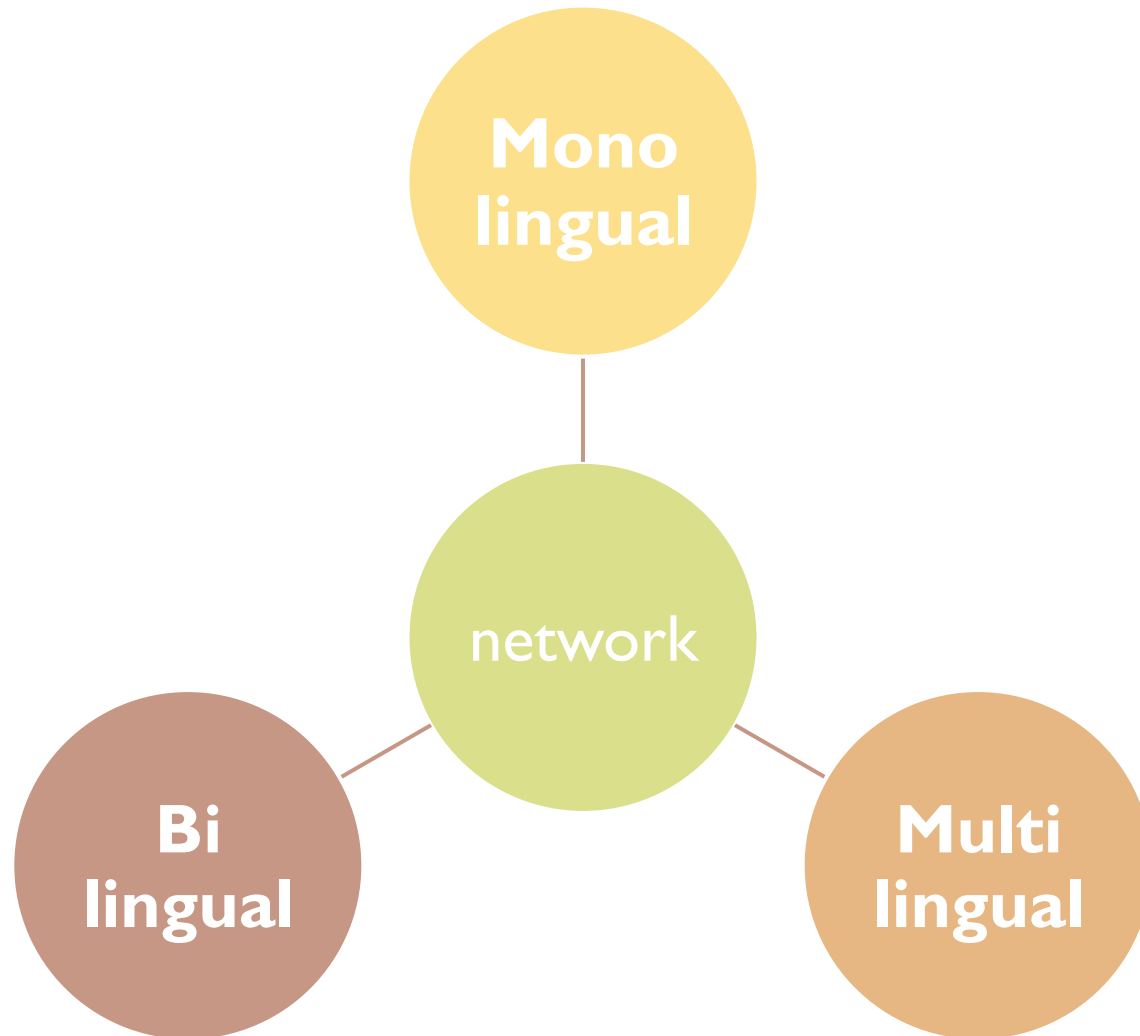
TECHNOLOGIC

- ▶ editorial, processing & publication tools
- ▶ XML-RDF configuration
- ▶ QA & statistics
- ▶ data maintenance, update & upgrade
- ▶ technical support
- ▶ digital applications
- ▶ R&D

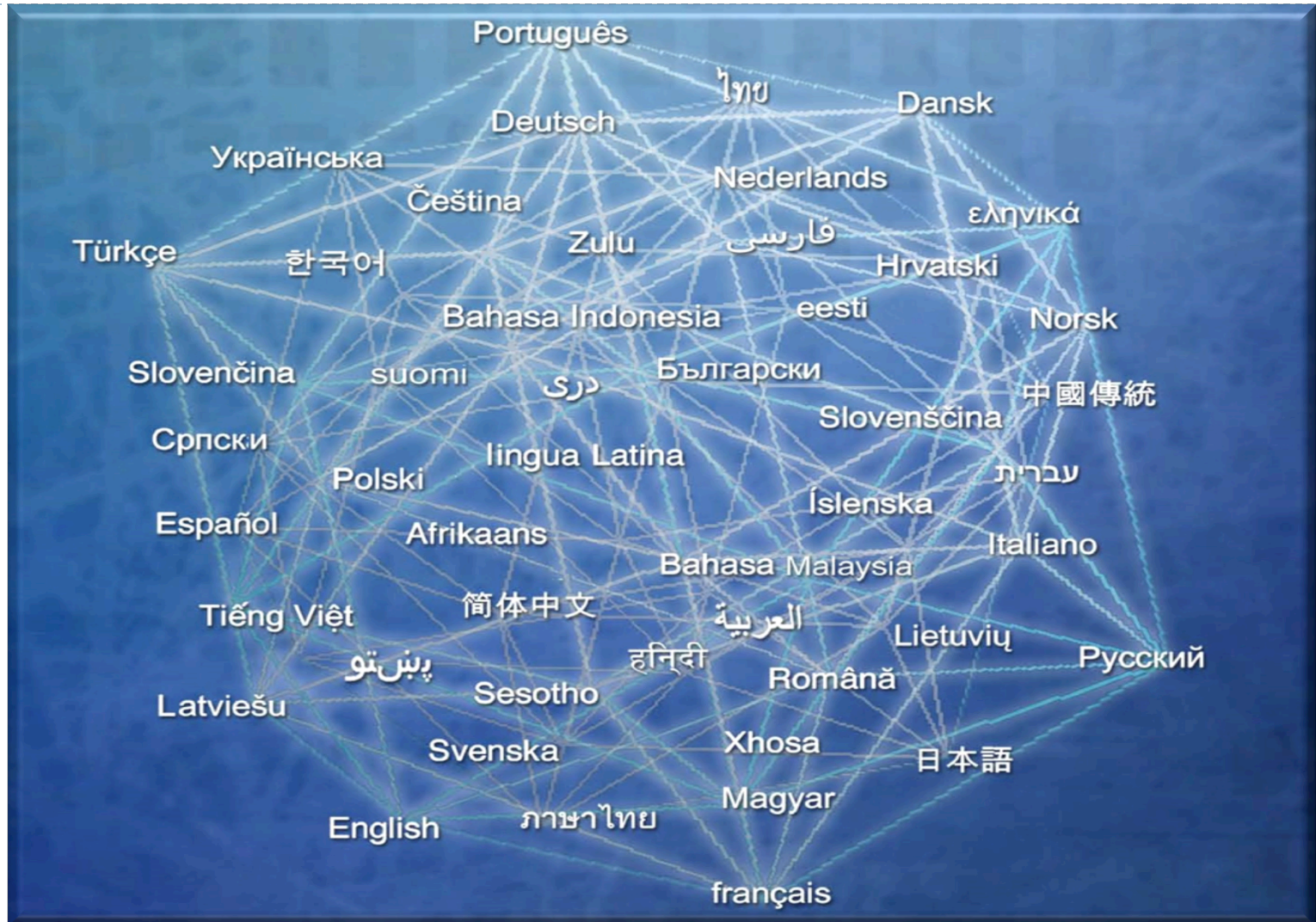
EVOLUTION

1. monolingual English learner's dictionary
2. semi-bilingual English learner's dictionary
3. (semi-)multilingual English dictionary
4. L2-English reversed indexes
5. L2, L3 etc. multilingual dictionaries
6. L2-L3 bilingual glossaries
7. multi-language networks

MULTI-LAYER



VISION



ENGLISH MULTILINGUAL

▶ PASSWORD semi-bilingual dictionary

▶ KEMD (44 languages)

Afrikaans | Arabic | Bulgarian | Catalan | Chinese
(Simplified | Traditional) | Croatian | Czech | Danish |
Dutch | English | Estonian | Farsi | Finnish | French |
German | Greek | Hebrew | Hindi | Hungarian | Icelandic |
Indonesian | Italian | Japanese | Korean | Latvian |
Lithuanian | Malay | Norwegian | Polish | Portuguese
(Brazil | Portugal) | Romanian | Russian | Serbian | Slovak |
Slovene | Spanish | Swedish |
Thai | Turkish | Ukrainian | Urdu | Vietnamese

L2 MULTILINGUALS

- ▶ Extract list of Translations of any language (L2) with their corresponding English (EN) Entries & POS
- ▶ Edit the L2 Translations into L2 Headwords, keeping the default EN links
- ▶ Revise the links from the new Headword & POS to the relevant sense of the EN Entry
- ▶ Each sense of the L2 Headword now addresses its counterpart sense(s) in the EN Entries, and through it translation equivalents in all other languages
- ▶ [Expand the lexical data of the L2 Headword and turn it into a full Entry]

DATA STRUCTURE

Main tables used for L2 Index generation

1. English HW table
2. Senses table
3. Translation table
4. L2 HW table
(used in L2 Index table, generated from the English HW, Senses and Translation tables)
5. L2 Senses table
(used for Tree and HTML preview, with English Words, Definitions and Examples tables)

PROCESS

- ▶ Generating an L2-English Index automatically
 - produce L2 Index table
 - produce EN Senses table
- ▶ Editing the L2 Index
 - include/exclude HW in L2 Index
 - revise the L2 HW and POS
 - add new L2 HW
 - revise the Senses – add, remove, re-order
- ▶ Translating multilingually
 - link L2 HW via EN Sense to all the translations

KIET. MAIN SCREEN

File Edit View Options Help

Index language: **French.** fr - French Update

Search: ban

| V | Headword | H | POS | HWordSrc |
|-------------------------------------|----------------|---|-------------|---------------|
| <input checked="" type="checkbox"/> | cadenasser | | verb | cadenasser |
| <input checked="" type="checkbox"/> | cadence | | noun | rythme, cad |
| <input checked="" type="checkbox"/> | cadet | | noun, ad... | cadet/-ette, |
| <input checked="" type="checkbox"/> | cadran | | noun | cadran |
| <input checked="" type="checkbox"/> | cadran solaire | | noun | cadran solai |
| <input checked="" type="checkbox"/> | cadre | | noun | cadre |
| <input checked="" type="checkbox"/> | cadrer (avec) | | verb | cadrer avec |
| <input checked="" type="checkbox"/> | cafard | | noun | cafard |
| <input type="checkbox"/> | cafard | | noun plural | cafard |
| <input checked="" type="checkbox"/> | cafarder | | | cafarder, rap |
| <input type="checkbox"/> | café | | adjective | (couleur) ca |
| <input checked="" type="checkbox"/> | café | | noun | café |
| <input checked="" type="checkbox"/> | caféine | | noun | caféine |

Senses

- frame** (noun)
 - something made to enclose something
- executive** (noun)
 - a person or body of people in an organization etc
- setting** (noun)
 - a background

Source language entry

cadre *noun*

- 1. frame** *noun*
something made to enclose something
◊ a picture frame = a window frame.
- 2. executive** *noun*
a person or body of people in an organization etc that has power to direct or manage
◊ He is an executive in an insurance company.
- 3. setting** *noun*
a background
◊ This castle is the perfect setting for a murder.

Exit

KIET. EDIT L2-ENGLISH INDEX (FRENCH)

Search in:

English Headwords Headword
 English Definitions

Headword:

POS:

Type the text to search (in the appropriate language):

| flag | HWord | POS | Definition |
|-------------------------------------|---------------|------|--|
| <input type="checkbox"/> | frame of mind | noun | mental state |
| <input type="checkbox"/> | frame | noun | the human body |
| <input checked="" type="checkbox"/> | frame | noun | something made to enclose something |
| <input type="checkbox"/> | frame | noun | a hard main structure round which something is built |
| <input type="checkbox"/> | frame | verb | to arrange false evidence so as to mislead |
| <input type="checkbox"/> | frame | verb | to act as a frame for |
| <input type="checkbox"/> | frame | verb | to put a frame around |

cadre (noun)

- frame (noun)**
 - something made to enclose something
◇ a picture frame = a window frame.
- executive (noun)**
 - a person or body of people in an organization etc that has power to direct or manage
◇ He is an executive in an insurance company.
- setting (noun)**
 - a background
◇ This castle is the perfect setting for a murder.

KIET. EDIT BY DEFINITION

The screenshot shows the 'Edit' window in KIET. The search criteria are: English Headwords (checked), English Definitions (checked), Headword: 'cadre', and POS: 'noun'. The search text is 'setting'. The results are displayed in a table and a tree view.

Search in:
 English Headwords Headword
 English Definitions

Headword: cadre
POS: noun
Edit

Type the text to search (in the appropriate language):
setting
Check all Restore to original

| flag | HWord | POS | Definition |
|-------------------------------------|------------------|------------|---|
| <input type="checkbox"/> | setting lotion | noun | a lotion that is used in setting the hair |
| <input checked="" type="checkbox"/> | setting | noun | music composed for a poem etc |
| <input type="checkbox"/> | setting | noun | an arrangement of jewels in eg a ring |
| <input checked="" type="checkbox"/> | setting | noun | a background |
| <input type="checkbox"/> | arson | noun | the crime of setting fire to (a building) |
| <input type="checkbox"/> | incendiary | adjecti... | used for setting (a building etc) on fire |
| <input type="checkbox"/> | insurance policy | noun | (a document setting out) an agreement |
| <input type="checkbox"/> | mass | noun | a setting to music of some of the words |
| <input type="checkbox"/> | scene | noun | the setting or background for a play |
| <input type="checkbox"/> | setting lotion | noun | a lotion that is used in setting the hair |
| <input type="checkbox"/> | sunset | noun | the setting of the sun, or the time of f |

cadre (noun)

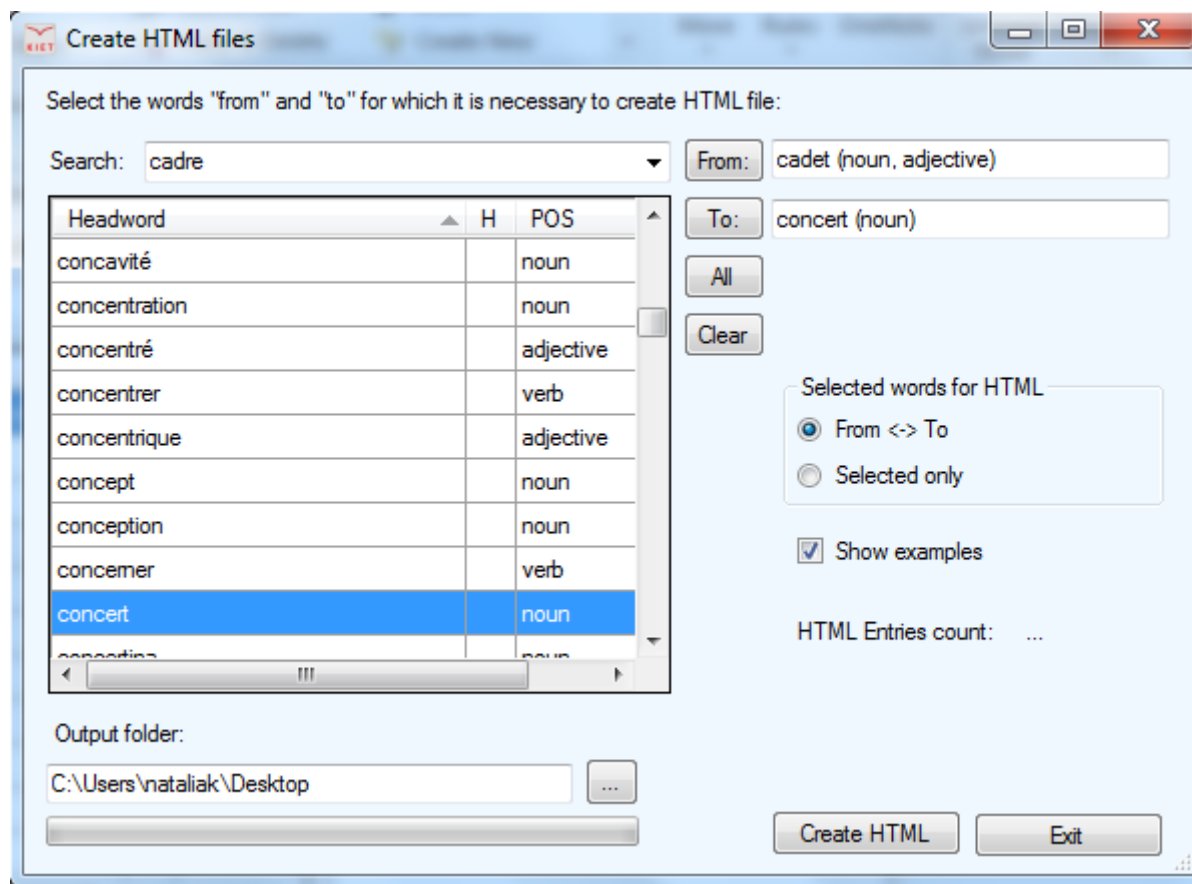
- executive (noun)
 - a person or body of people in an organization etc that has power to direct or manage
He is an executive in an insurance company.
- frame (noun)
 - something made to enclose something
a picture frame = a window frame.
- setting (noun)
 - a background
This castle is the perfect setting for a murder.
 - setting (noun)
 - music composed for a poem etc
settings of folk songs.

cadre noun

- executive noun**
a person or body of people in an organization etc that has power to direct or manage
He is an executive in an insurance company.
- frame noun**
something made to enclose something
a picture frame = a window frame.
- setting noun**
a background
This castle is the perfect setting for a murder.
- setting noun**
music composed for a poem etc
settings of folk songs.

Cancel Save

KIET. EXPORT TO HTML



SAMPLE. GERMAN-ENGLISH INDEX

messen *verb*

1. **gauge** to measure (something) very accurately
2. **measure** to find the size, amount etc of (sth)
3. **measure** to show the size, amount etc of
4. **measure** (with **against**, **besides** etc) to judge in comparison with
5. **measure** to be a certain size
6. **meter** to measure (*especially* electricity etc) by using a meter
7. **take** to make a note, record etc

SAMPLE. GERMAN MULTILINGUAL (1)

messen *verb*

I. to measure (something) very accurately

af meet | **ar** ي | **bg** измервам точно | **br** medir | **ca** mesurar, calibrar | **cs** (z)měřit | **dk** måle | **el** (κατα)μετρώ με ακρίβεια | **en** gauge | **es** medir, calibrar | **et** mõõtma | **fa** با دقت اندازه گیری | **fi** mitata | **fr** mesurer, jauger | **he** תִּימָן | **hi** प्रमाण, आयाम | **hr** mjeriti | **hu** megmér | **id** mengukur | **is** mæla | **it** calcolare | **ja** 測る | **ko** 정확히 측정하다 | **lt** matuoti | **lv** mērīt | **ml** mengukur | **nl** meten | **no** måle (opp) | **pl** wymierzyć | **pt** medir | **ro** a măsură | **ru** измерять | **sk** odmerať | **sl** izmeriti | **sr** izmeriti | **sv** mäta | **th** วัดด้วยมาตรวัด; เครื่องวัด | **tr** ölçmek | **tw** 精確測量 | **uk** виміряти | **ur** کسی چیز کو ناپنا | **vi** đo | **zh** 精确测量

SAMPLE. GERMAN MULTILINGUAL (2)

messen *verb*

2. to find the size, amount etc of (something)

af meet | ar ي | bg измервам | br medir | ca mesurar |
cs (z)měřit | dk måle | el μετρώ | en measure | es medir |
et mõõtma | fa اندازه گیری کردن | fi mitata | fr mesurer | he תימדל
| hi नापना | hr mjeriti | hu (meg)mér | id mengukur | is mæla |
it misurare | ja 測る | ko 치수를 재다 | lt (iš)matuoti | lv no |
ml mengukur | nl meten | no måle, ta mål av | pl (wy)mierzyć |
pt medir | ro a măsură | ru измерять | sk odmerať | sl izmeriti | sr
izmeriti | sv mäta | thวัดขนาด (ความยาว, ความสูง, ความเร็ว
 ฯลฯ) |
tr ölçmek | tw 測量 | uk міряти, вимірювати |
ur مقدار، حجم وغیرہ معلوم کرنا | vi đo lường | zh 測量

GLOBAL SERIES

- ▶ Arabic
- ▶ Chinese Simp.
- ▶ Chinese Trad.
- ▶ Czech
- ▶ Danish
- ▶ Dutch (2)
- ▶ English
- ▶ French (2)
- ▶ German (2)
- ▶ Greek
- ▶ Hebrew
- ▶ Italian (2)
- ▶ Japanese
- ▶ Korean
- ▶ Latin
- ▶ Norwegian
- ▶ Polish
- ▶ Portuguese Br.
- ▶ Portuguese Pt.
- ▶ Russian
- ▶ Spanish (3)
- ▶ Swedish (2)
- ▶ Thai
- ▶ Turkish

THANK YOU

[θæŋk ju:] *interj.* I thank you: *Thank you for your attention!*

Afrikaans **dankie**

Arabic ش

Bulgarian **благодаря**

Chinese Simplified 谢谢(你)

Chinese Traditional 謝謝(你)

Croatian **hvala**

Czech **děkuji**

Danish **tak**

Dutch **dank je**

Estonian **aitäh, tänan teid**

Farsi ممنون

Finnish **kiitos**

French **merci**

German **danke**

Greek (σε, σας) ευχαριστώ

Hebrew תודה

Hindi धन्यवाद देने या मना करने का एक

Hungarian **köszönöm!**

Icelandic **þakka þér**

Indonesian **terima kasih**

Italian **grazie**

Japanese ありがとう

Korean 감사합니다

Latvian **paldies; pateicos**

Lithuanian **ačiū**

Malay **terima kasih**

Norwegian **tusen takk (for)**

Polish **dziękuję**

Portuguese Brazil **obrigado/-da**

Portuguese Portugal **obrigado/-da**

Romanian **mulțumesc**

Russian **благодарю**

Serbian **hvala**

Slovak **d'akujem**

Slovene **hvala**

Spanish **gracias**

Swedish **tack [ska du/ni ha]!, tackar!**

Thai การแสดงความขอบคุณ

Turkish **teşekkür ederim**

Ukrainian **дякую; спасиби**

Urdu آپ کا شکریہ

Vietnamese **cảm ơn**