Linked Data for Language Technology
roadmapping workshop
Athens, 21 March 2014

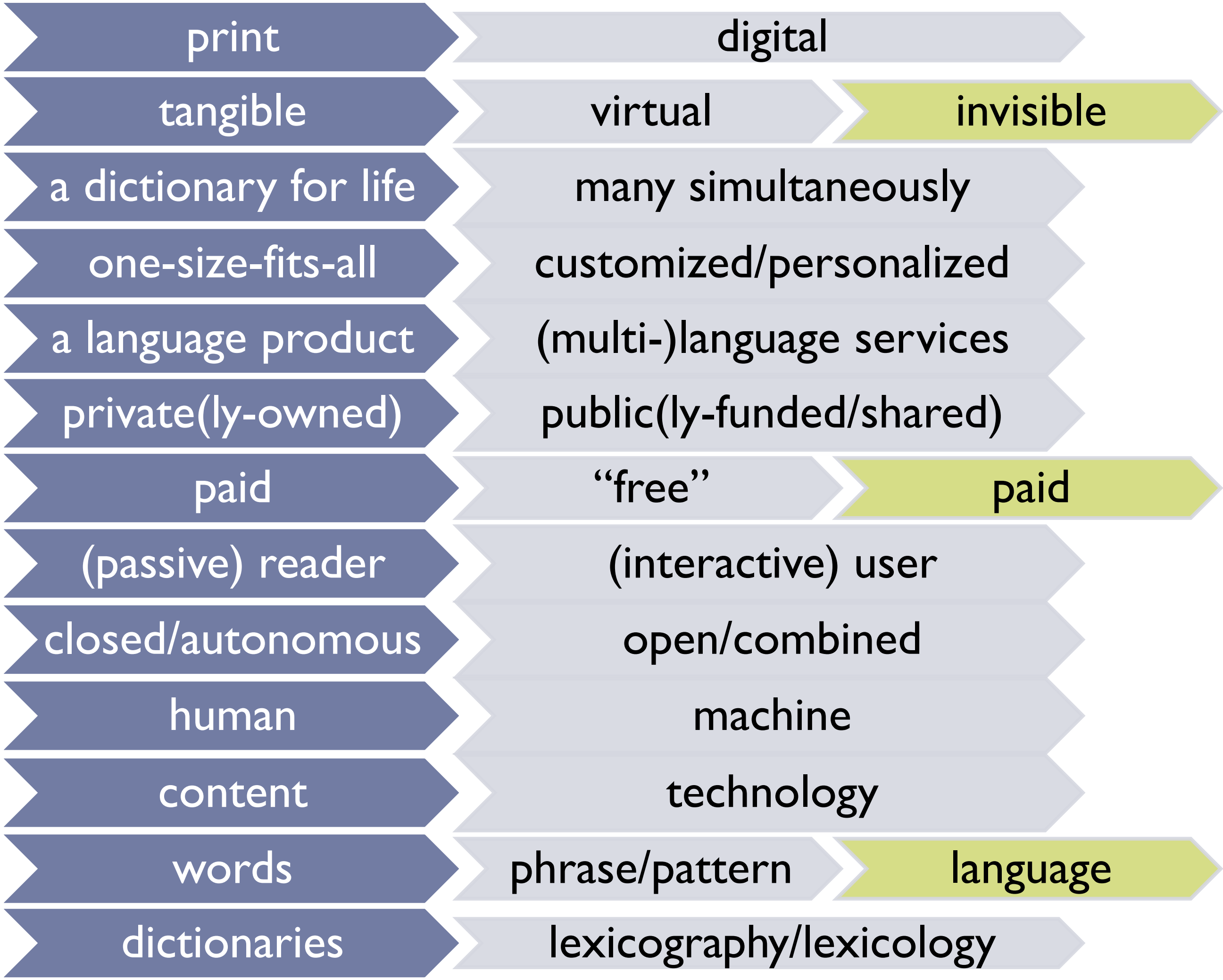
Developing multi-language dictionary data for NLP use

Ilan Kernerman
K Dictionaries, Tel Aviv

highlights

- Trends & KD
- English multilingual dictionary
  → semi-automatically generated multilingual glossaries
- Global series
  monolingual ↔ bilingual ↔ multilingual
  ↕
  multi-layer
- Lexicography & LT

trends

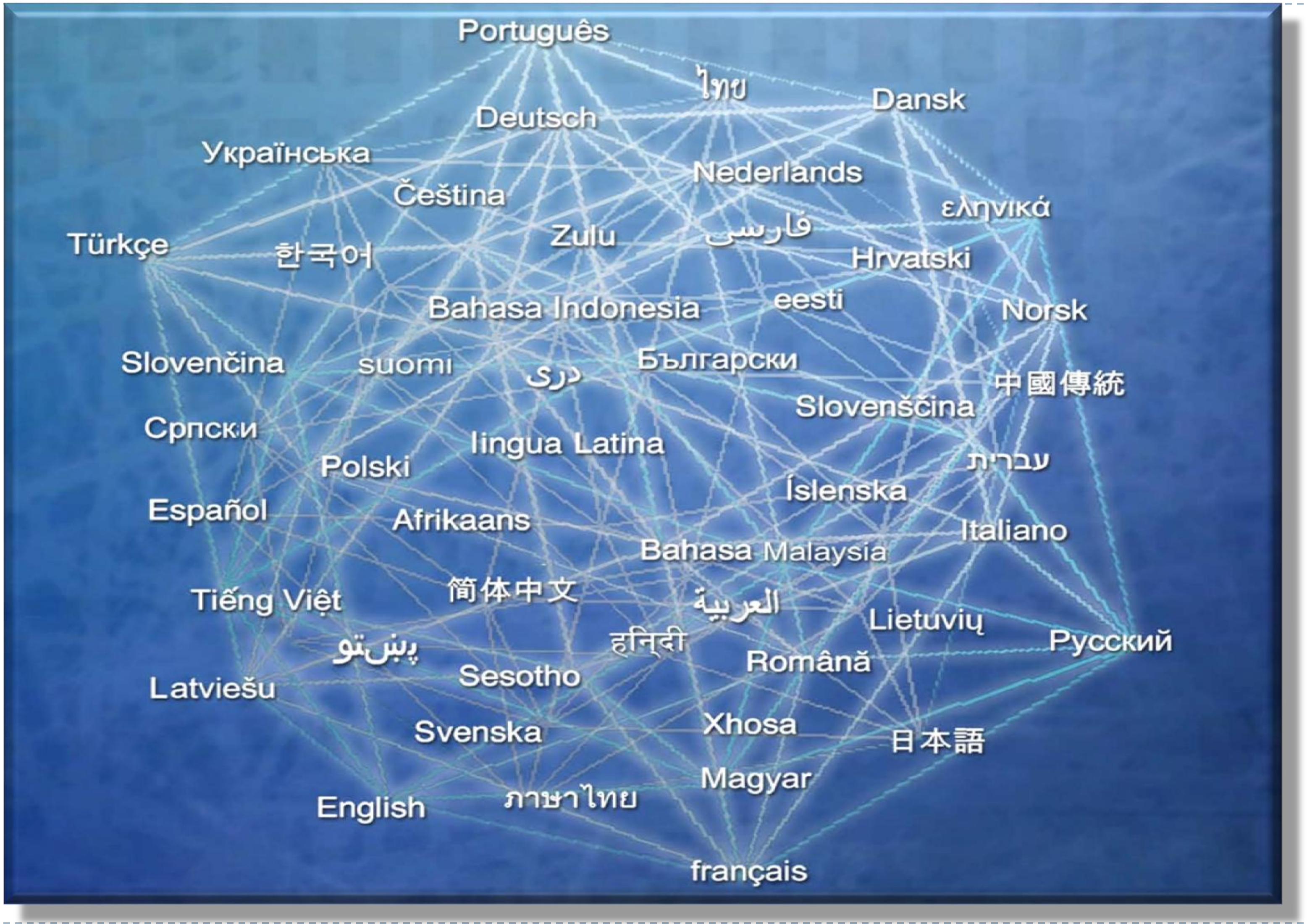


print	digital	
tangible	virtual	invisible
a dictionary for life	many simultaneously	
one-size-fits-all	customized/personalized	
a language product	(multi-)language services	
private(ly-owned)	public(ly-funded/shared)	
paid	“free”	paid
(passive) reader	(interactive) user	
closed/autonomous	open/combined	
human	machine	
content	technology	
words	phrase/pattern	language
dictionaries	lexicography/lexicology	

kdictionaries – *overview*

- Established in 1993
- Based in Tel Aviv, active globally
- Creates lexicographic content, applications and tools
- Develops monolingual, bilingual and multilingual data in 40+ languages
- Publishes annual newsletter, as well as collections of papers on dictionaries and lexicography
- Collaborates with lexicographers and translators, software engineers and designers, publishers and ICT firms, the academe and professional associations

kdictionaries – *foresight*



English multilingual

- The semi-bilingual English learner's dictionary
  http://kdictionaries-online.com/Password.aspx?Dictionary=14
- GlobalDix – 2001 Kielikone, 21 languages
  http://kdictionaries.com/kdn/kdn9-4.html
- KEMD (46 languages)
  Afrikaans | Arabic | Bulgarian | Catalan | Chinese (Simplified | Traditional) | Croatian | Czech |Danish | Dari | Dutch | English | Estonian | Farsi | Finnish | French | German | Greek | Hebrew | Hindi | Hungarian | Icelandic | Indonesian | Italian | Japanese | Korean | Latvian | Lithuanian | Malay | Norwegian | Pashtu | Polish | Portuguese (Brazil | Portugal) | Romanian | Russian | Serbian | Slovak | Slovene | Spanish | Swedish | Thai | Turkish | Ukrainian | Urdu | Vietnamese

12 multilingual glossaries

- Reverse the translations from any language (L2) back to English (EN)
- Edit the L2 Translations into L2 Headwords, maintaining the default links to the EN Entries
- Edit the links from the L2 Headword to the original specific meaning of the EN Entry
- Each meaning of the L2 Headword refers to appropriate meanings of the EN Entry – and thereby to all languages
- Expand the lexical data concerning the L2 Headword and turn it into a full L2 Entry

Swedish sample – *bortsprungen* (1)

1. **runaway** *noun*

   a person, animal *etc* that runs away

   ◊ *The police caught the two runaways.*

   ▫ *(also adjective) a runaway horse.*

af - weglоper | ar - هارِب، شارِد، جامِح | bg - беглец | br - fugitivo | ca - fugitiu | cs - uprchlík/-ice, uprchlý | de - der/die Ausreißer(in); durchgebrannt | dk - bortløben | el - φυγάδας | es - fugitivo | et - põgenik | fa - فراری | fi - karkuri | fr - fugitif/-ive | he - בּוֹרֵחַ | hi - अनियंत्रित, उच्छृंखल, बहुत सहज | hr - odbjegao | hu - szökevény | id - pelarian | it - fuggiasco, fuggitivo | ja - 逃亡者 | ko - 도망자 | lt - pabėgėlis; pabėgęs | lv - bēglis; izbēdzis | ml - cabut lari | nl - vluchteling | pl - zbiec | prs - فراری | ps - فراری | pt - fugitivo | ro - evadat, fugar | ru - беглец | sk – utečenec/-ka; na úteku, ktorý ušiel | sl - ubežnik; pobegel | sr - odbegao | th - ผู้หลบหนี | tr - kaçak, firari | tw - 逃跑的人或動物 | uk - утікач; дезертир | ur - فرار ہو جانا | vi - kẻ chạy trốn | zh - 潜逃者，逃跑者

Swedish sample – *bortsprungen* (2)

2. stray *adjective*

wandering or lost

◊ *stray cats and dogs.*

af - weglopend | ar - شارِد، ضال، تائِه | bg - изгубен | br - perdido | ca - perdut, extraviat, llista de carrers | cs - zatoulaný | de - streunend | dk - omstrejfende; herreløs | el - αδέσποτος | es - perdido, extraviado, callejero | et - hulkuv | fa - گمشده | fi - kuljeksiva | fr - errant | he - חֲסַר בַּיִת | hi - भटका, भूलाभटका | hr - zalutao, zabludio | hu - elkóborolt | id - sesat | it - randagio | ja - はぐれた | ko - 길 잃은 | lt - benamis, valkataujantis | lv - noklīdis; klaiņojošs | ml - terbiar | nl - zwerf- | pl - bezdomny | prs - گمشده | ps - ورک شوی | pt - perdido | ro - rătăcit | ru - бездомный | sk - zatúlaný | sl - klateški | sr - izgubljen | th - ซึ่งพลัดหลง | tr - başıboş dolaşan | tw - 漫遊的 | uk - бездомний | ur - آواره یا لاوارث | vi - lạc, mất | zh - 漫游的

global series – *overview*

- A rich lexicographic dataset for each language
- Based on English pedagogical lexicography principles
- Lexical deconstruction and reconstructions
- Serve as a base to develop monolingual, bilingual and multilingual solutions
- Add full translation equivalents to L1 in other languages
- Adapt content specifically for each language pair
- Use the data differently to suit each target group and usage (L1/L2, language learning, translation, etc.)
- Apply the data in electronic forms
- Incorporate the data with NLP/LT

global series – *languages*

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

global series – *linguistic DNA*

- AlternativeScripting
- AlternativeSpelling
- Antonym
- CompositionalPhrase
- CrossReference
- Definition
- Example
- GeographicalUsage
- GrammaticalGender
- GrammaticalNumber
- HomographNumber
- Lemma
- Morphology
- PartOfSpeech
- Pronunciation
- RangeOfApplication
- Register
- SenseIndicator
- SenseQualifier
- SubCategorization
- SubjectField
- Synonym

global series – *microstructure*

headword

- Lemma
- Pronunciation
- Part of speech
- Gender
- Grammatical number
- Irregular forms
- Alternative scripts

global series – *microstructure*

attributes

- Geographical Usage
- Subject Field
- Register
- Range of Application
- Sense Qualifier
- Synonym
- Antonym

global series – *microstructure*

definition

- A succinct definition for each sense of the entry

sense indicator

- Hypernym or context ‘preposition-filler’ (e.g. *of* house repairs) for sense disambiguation in polysemous entries

translation

- L2 equivalent for each sense, with pronunciation, conjugations, grammatical gender and number (NOT a translation of the definition itself)

global series – *indicator vs definition*

definition

- ACCUEIL *nm*
  1. manière de recevoir qqn ou qqch ◊ *faire bon / mauvais accueil à qqn*
  2. lieu ou on recoit des visiteurs ◊ *Adressez-vous à l'accueil !*

sense indicator

- ACCUEIL *nm*
  1. réception ◊ *faire bon / mauvais accueil à qqn*
  2. lieu ◊ *Adressez-vous à l'accueil !*

global series – *microstructure*

example of usage

- Example(s) of usage for each sense of polysemous entries, preferably consisting of a short phrase (rather than a full sentence)

translation

- L2 idiomatic equivalent of the example

global series – *microstructure*

phrases

- Idiomatic expressions (idioms, collocations, phrasal verbs, etc.) with/without definition and/or example, as part of a given sense or consisting of a sense on its own

translation

- L2 idiomatic equivalent of the phrase (NOT of its definition) and its examples

global series – *microstructure*

sub-headword

- Run-on of the main entry, may include any of its components

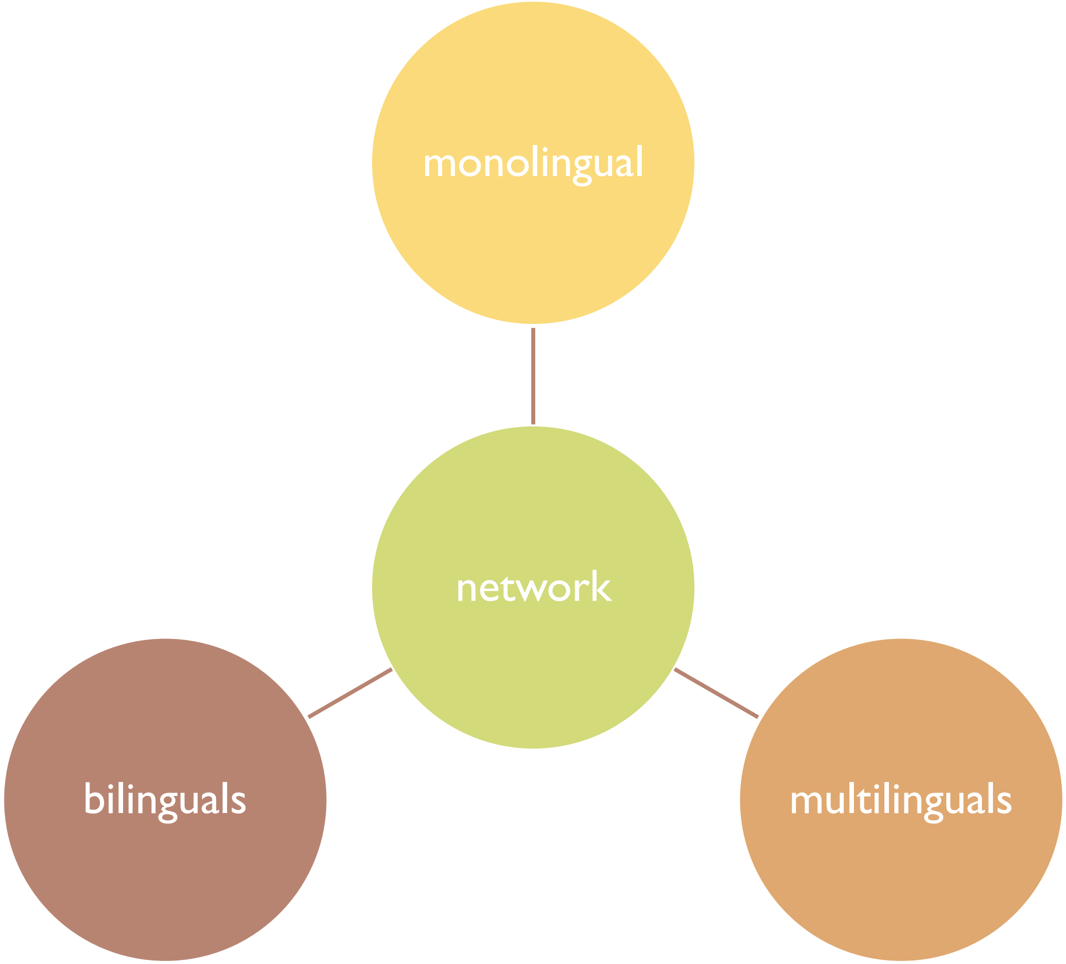
translation

- L2 equivalent of each component

global series - *sample*

- monolingual
  http://kdictionaries-online.com/nIMLDSplus.aspx
- bilingual
- multilingual

global series – *multi-layer creation & processing*



global series – *multi-layer*

- http://kdictionaries-online.com/frMLDS.aspx?Languages=ar,zh,nl,de

LT application

- Machine translation
- Language learning
- Text processing, search engines, etc.
- User guides, travel guides, menus, etc.
- Text-To-Speech, STT, etc.

thank you

- THANK YOU [ˈθæŋ juː] *interjection*
  an expression of thanks
  ◊ *Thank you for your attention* ☺

Ilan Kernerman
K Dictionaries Ltd
Nahum 8 Tel Aviv
+972 3 5468102
ilan@kdictionaries.com

KDICTIONARIES