Toward self-sustaining community-driven online terminology development

C. Maria Keet[1,2] and Graham Barbour[1]

[1] School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, South Africa
[2] UKZN/CSIR-Meraka Centre for Artificial Intelligence Research, South Africa
{keet,barbour}@ukzn.ac.za

7th Annual Teaching & Learning Higher Education Conference 2013
Pinetown, South Africa, September 25-27, 2013

Outline

1. Introduction
2. Related works
3. First results
   - Workshop experiment
   - Commuterm crowdsourcing tool

Outline

1. Introduction
2. Related works
3. First results
   - Workshop experiment
   - Commuterm crowdsourcing tool

Background

- Principal obstacle for teaching and tutoring computing and information technology (CS&IT) in isiZulu is the lack of isiZulu language terminology on the whole, and fragmented knowledge of the existing isiZulu words in the fields of CS&IT even among isiZulu speakers
- Other language areas (e.g., German, Spanish, Italian): gradual development over past 70 years, national bodies enforcing new terms
  - E.g., in 2013: Académie Francaise instituted *mot-dièse* for 'hashtag' and Real Academia Española instituted *whatsappear* as verb for using the WhatsApp application
- For isiZulu: Google's localization, using, e.g., *isilungiselelo* for settings (fine) and *idrayivu* for the Google drive (not fine)

Background

- Principal obstacle for teaching and tutoring computing and information technology (CS&IT) in isiZulu is the lack of isiZulu language terminology on the whole, and fragmented knowledge of the existing isiZulu words in the fields of CS&IT even among isiZulu speakers
- Other language areas (e.g., German, Spanish, Italian): gradual development over past 70 years, national bodies enforcing new terms
  - E.g., in 2013: Académie Française instituted *mot-dièse* for 'hashtag' and Real Academia Española instituted *whatsappear* as verb for using the WhatsApp application
- For isiZulu: Google's localization, using, e.g., *isilungiselelo* for settings (fine) and *idrayivu* for the Google drive (not fine)

Background

- Principal obstacle for teaching and tutoring computing and information technology (CS&IT) in isiZulu is the lack of isiZulu language terminology on the whole, and fragmented knowledge of the existing isiZulu words in the fields of CS&IT even among isiZulu speakers
- Other language areas (e.g., German, Spanish, Italian): gradual development over past 70 years, national bodies enforcing new terms
    - E.g., in 2013: Académie Française instituted *mot-dièse* for 'hashtag' and Real Academia Española instituted *whatsappear* as verb for using the WhatsApp application
- For isiZulu: Google's localization, using, e.g., *isilungiselelo* for settings (fine) and *idrayivu* for the Google drive (not fine)

Terminology development

- Terminologies, such as thesauri and structured controlled vocabularies, and broader use, including glossaries
- Pure linguistic approach; e.g., the "conceptual blending" (compounding) for creating new isiZulu terms (Buthelezi (2008))
    - Creating new words by combining existing ones
    - E.g., for CS: 'programming' as *ukwakhuhlelo*, contracting *ukwakha* ('to build') and *uhlelo* ('arrangement' or 'grammar')
- Traditional/typical approach: time and resource-consuming workshops with stakeholders; e.g.:
    - Stellenbosch University for isiXhosa: but not CS&IT, and for payment
    - Department of Arts and Culture of South Africa (2005): 135 terms, in the 11 official languages in SA: 76 reasonably within IT computer literacy

Terminology development

- Terminologies, such as thesauri and structured controlled vocabularies, and broader use, including glossaries
- Pure linguistic approach; e.g., the “conceptual blending” (compounding) for creating new isiZulu terms (Buthelezi (2008))
  - Creating new words by combining existing ones
  - E.g., for CS: ‘programming’ as *ukwakhuhlelo*, contracting *ukwakha* (‘to build’) and *uhlelo* (‘arrangement’ or ‘grammar’)
- Traditional/typical approach: time and resource-consuming workshops with stakeholders; e.g.:
  - Stellenbosch University for isiXhosa: but not CS&IT, and for payment
  - Department of Arts and Culture of South Africa (2005): 135 terms, in the 11 official languages in SA: 76 reasonably within IT computer literacy

Terminology development

- Terminologies, such as thesauri and structured controlled vocabularies, and broader use, including glossaries
- Pure linguistic approach; e.g., the “conceptual blending” (compounding) for creating new isiZulu terms (Buthelezi (2008))
  - Creating new words by combining existing ones
  - E.g., for CS: ‘programming’ as *ukwakhuhlelo*, contracting *ukwakha* (‘to build’) and *uhlelo* (‘arrangement’ or ‘grammar’)
- Traditional/typical approach: time and resource-consuming workshops with stakeholders; e.g.:
  - Stellenbosch University for isiXhosa: but not CS&IT, and for payment
  - Department of Arts and Culture of South Africa (2005): 135 terms, in the 11 official languages in SA: 76 reasonably within IT computer literacy

Setting

- Historically, politically, and economically, it is urgent to develop scientific terminology
- Has to occur in a much shorter timespan than occurred with some other languages

⇒ How to achieve rapid terminology development?

⇒ In a manner that terminology development is by the people for the people

Setting

- Historically, politically, and economically, it is urgent to develop scientific terminology
- Has to occur in a much shorter timespan than occurred with some other languages

⇒ How to achieve rapid terminology development?

⇒ In a manner that terminology development is by the people for the people

Proposal

- Rapid terminology development using “games with a purpose”
- Crowdsourcing to obtain input from a large group of isiZulu speakers
- Test the new method with development of a computer science terminology in isiZulu
- Verify the method in another subject domain
- Generalise the new method to any language, any domain

Proposal

- Rapid terminology development using “games with a purpose”
- Crowdsourcing to obtain input from a large group of isiZulu speakers
- Test the new method with development of a computer science terminology in isiZulu
- Verify the method in another subject domain
- Generalise the new method to any language, any domain

Outline

1. Introduction
2. Related works
3. First results
   - Workshop experiment
   - Commuterm crowdsourcing tool

Crowdsourcing

- Crowdsourcing: a wordplay outsourcing where there is another "pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R & D" (Howe (2006))
- Now part of a general-purpose problem solving method of mass collaboration systems on the Web (Doan et al. (2011))
  - Early examples: SETI@home (Korpela et al. (2001)) and ESP game (Ahn & Dabbish (2004))
- Wide range of tasks: from protein folding to collaborative and distributed algorithm development
- One language-related: DuoLingo (`http://www.duolingo.com`) for translations of major languages
- None on crowdsourcing terminologies

Crowdsourcing

- Crowdsourcing: a wordplay outsourcing where there is another “pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R & D” (Howe (2006))
- Now part of a general-purpose problem solving method of mass collaboration systems on the Web (Doan et al. (2011))
    - Early examples: SETI@home (Korpela et al. (2001)) and ESP game (Ahn & Dabbish (2004))
- Wide range of tasks: from protein folding to collaborative and distributed algorithm development
- One language-related: DuoLingo (`http://www.duolingo.com`) for translations of major languages
- None on crowdsourcing terminologies

Crowdsourcing

- Crowdsourcing: a wordplay outsourcing where there is another "pool of cheap labor: everyday people using their spare cycles to create content, solve problems, even do corporate R & D" (Howe (2006))
- Now part of a general-purpose problem solving method of mass collaboration systems on the Web (Doan et al. (2011))
    - Early examples: SETI@home (Korpela et al. (2001)) and ESP game (Ahn & Dabbish (2004))
- Wide range of tasks: from protein folding to collaborative and distributed algorithm development
- One language-related: DuoLingo (`http://www.duolingo.com`) for translations of major languages
- None on crowdsourcing terminologies

Crowdsourcing and mass collaboration

- Different types of ‘mass labour’ online, based on, a.o. (Good & Su (2013)):
    - Volunteer labour (ESP game) vs. forced labour (ReCAPTCHA)
    - Microtasks (gene and photo annotations) vs. macrotasks (protein folding with FoldIT!)
    - For ‘fun’, payment by game, payment for best solution
- Requires different design choices regarding recruitment, retention, evaluation user contributions, calculations on solution of task
- How to handle malicious users

Crowdsourcing and mass collaboration

- Different types of 'mass labour' online, based on, a.o. (Good & Su (2013)):
  - Volunteer labour (ESP game) vs. forced labour (ReCAPTCHA)
  - Microtasks (gene and photo annotations) vs. macrotasks (protein folding with FoldIT!)
  - For 'fun', payment by game, payment for best solution
- Requires different design choices regarding recruitment, retention, evaluation user contributions, calculations on solution of task
- How to handle malicious users

Outline

1 Introduction

2 Related works

3 First results

- Workshop experiment
- Commuterm crowdsourcing tool

Methodology

- State of the art: literature and a poll (tbd due to localization of survey software)
- Workshop experiment (completed)
- Crowdsourcing experiment for CS (launch within 2 weeks)
- Crowdsourcing experiment for another subject domain

Workshop experiment set up (summary)

- Participants: 10 students with isiZulu as home language (3rd year and honours students in computer science or computer science & information systems)
- Duration: 2 hours
- Incentives: the honour of being at the forefront of this endeavour, and pizza and softdrinks afterward
- For each term provided by the RAs, note term, consensus or not, synonyms

Results (snapshot)

- 15 students participated; 9 CS or IS honours students, and the remainder in their final year BSc CS; gender distribution: 5 female, 10 male
- Typically, meaning of the term was discussed before reaching an agreement on possible alternatives
- 37 entities in CS, focussed on programming and networking, beyond computer literacy
- Among others: *indlela yokwenza* for 'algorithm', *ukushintsha ufuzo* for 'overriding', *amalungu ohlelo ahlelekile* for 'formal parameter list'

Results (snapshot)

- 15 students participated; 9 CS or IS honours students, and the remainder in their final year BSc CS; gender distribution: 5 female, 10 male
- Typically, meaning of the term was discussed before reaching an agreement on possible alternatives
- 37 entities in CS, focussed on programming and networking, beyond computer literacy
- Among others: *indlela yokwenza* for 'algorithm', *ukushintsha ufuzo* for 'overriding', *amalungu ohlelo ahlelekile* for 'formal parameter list'

Workshop experiment

Discussion (snapshot)

- Overlap of 5 English terms with the Dept. of Arts & Culture (DAC) list, with an empty intersection
    - E.g.: database – *inqolobane* (our experiment) – *ulwazi olugciniwe, ulwazi olulondoloziwe, imininingo egciniwe* (DAC)
    - DAC wrong regarding the *ulwazi* (knowledge), because database ≠ knowledge base
- Real transformations of the underlying meaning, not simply zulufying English

Discussion (snapshot)

- Overlap of 5 English terms with the Dept. of Arts & Culture (DAC) list, with an empty intersection
  - E.g.: database – *inqolobane* (our experiment) – *ulwazi olugciniwe, ulwazi olulondoloziwe, imininingo egciniwe* (DAC)
  - DAC wrong regarding the *ulwazi* (knowledge), because database $\neq$ knowledge base
- Real transformations of the underlying meaning, not simply zulufying English

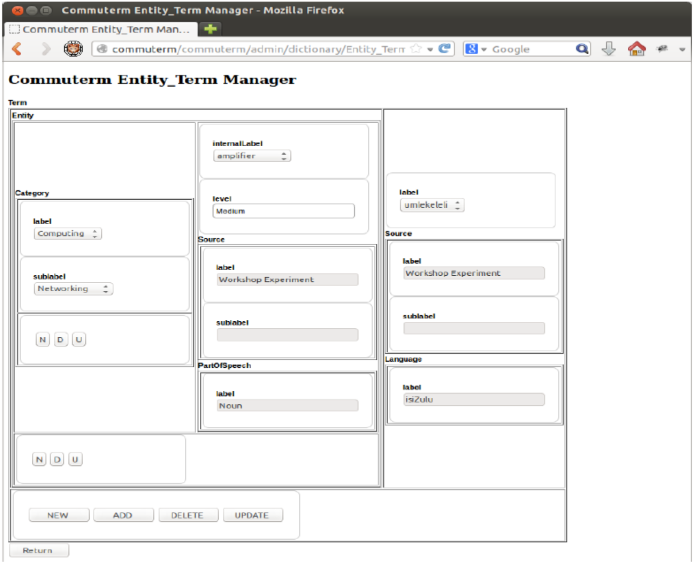
Design

- Software development methodology: roughly, an iterative version of the waterfall methodology
- Doan et al. (2011)'s 'four questions that all MC systems have to answer'
- Based on that, game requirements (using Good & Su (2013)'s categorisation) and system requirements engineering
- System development: database with web-based front-end
- Terminology list development, with definitions, source, level
- Implementation
- Currently in the testing phase

Demo

- Screenshots in this set of slides – demo at the presentation
- Note: this is an alpha version
- Note: the live version of the user interface of the game is in isiZulu (not English, as in the following screenshots); current screenshots just to give a general idea regarding functionality

Back-end: adding entities


Commuterm Entity_Term Manager - Mozilla Firefox
Commuterm Entity_Term Man...
commuterm/commuterm/admin/dictionary/Entity_Term
Google
N
D
U
N
D
U
NEW
ADD
DELETE
UPDATE
Return

Commuterm Entity_Term Manager

Term

Entity

Category

label: Computing

sublabel: Networking

internalLabel: amplifier

level: Medium

Source

label: Workshop Experiment

sublabel:

PartOfSpeech

label: Noun

label: umlekeleli

Source

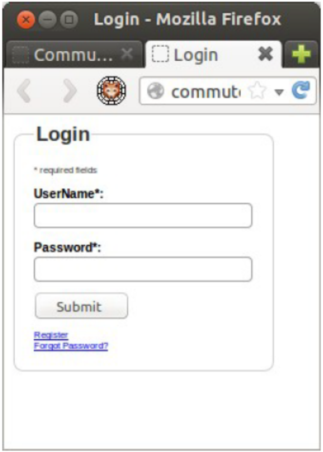
label: Workshop Experiment

sublabel:

Language

label: isiZulu

User login screen



Login - Mozilla Firefox

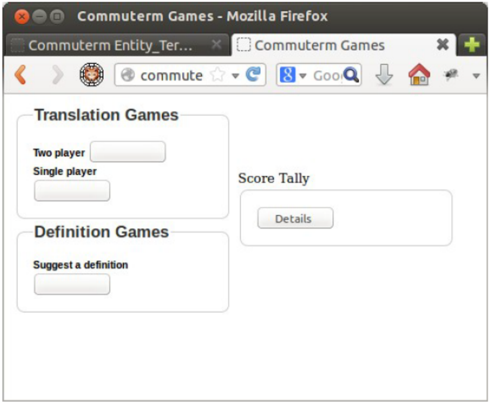
Login

* required fields

UserName*:

Password*:

Some user activity choices



Commuterm Games - Mozilla Firefox

Commuterm Entity_Ter...

Commuterm Games

commute

Translation Games

Two player

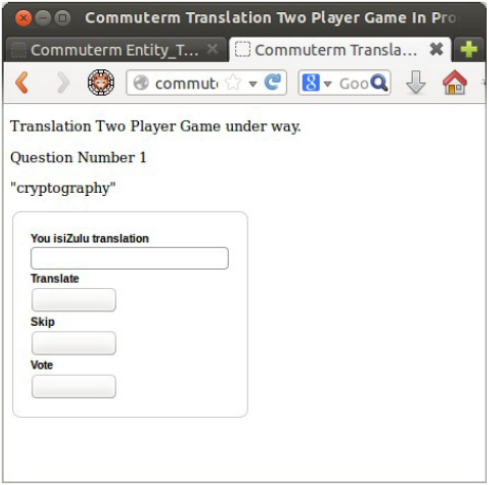
Single player

Definition Games

Suggest a definition

Score Tally

Snapshot of a game



Translation Two Player Game under way.

Question Number 1

"cryptography"

You isiZulu translation

References

Ahn, L. v. & Dabbish, L. (2004). Labeling images with a computer game. CHI Letters, 6, 319–326.

Buthelezi, T. M. (2008). Exploring the role of conceptual blending in developing the extension of terminology in isiZulu language. Alternation, 15, 181–200.

Department of Arts and Culture of South Africa (2005). Multilingual terminology for information communication technology. Tech. Rep. First edition, Department of Arts and Culture. http://www.dac.gov.za/chief_directorates/NLS/Website%20Multilingual%20ICTpdf%2019%20nov%202008.pdf.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. Communications of the ACM, 54, 86–96.

Good, B. M. & Su, A. I. (2013). Crowdsourcing for bioinformatics. Bioinformatics, 29, 1925–1933.

Howe, J. (2006). The rise of crowdsourcing. Wired, 14.06. http://www.wired.com/wired/archive/14.06/crowds.html. Last accessed: 25-9-2013.

Korpela, E., Werthimer, D., Anderson, D., Cobb, J., & Lebofsky, M. (2001). SETI@home – massively distributed computing for SETI. Computing in Science & Engineering, 3, 78–83.

Thank you!

Acknowledgements
Terminology workshop participants: Nelisiwe Dlamini, Sibonelo Dlamini, Bonginkosi Gina, Micah Khoza, Senzo Maphumulo, Snenhlanhla Mazibuko, Dumisani Mdlalose, Senzo Mthembu, Sthembile Mthembu, Sanele Ncwane, Zuzile Nxumalo, Banele Nzama, Zakhele Shoba, Tanita Singano, Welcome Sithole
Research assistants: Asive Dlaba, Siyabonga Madlala, Yanelang Motloung, Sibonginhlahla Nkosi
Funding: University Language Board, UKZN