Research Networking Programmes

Short Visit Grant ☒ or Exchange Visit Grant ☐

(please tick the relevant box)

Scientific Report

Scientific report (one single document in WORD or PDF file) should be submitted online within one month of the event. It should not exceed eight A4 pages.

Proposal Title: Openlexicons project

Application Reference N°: 5627

1) Purpose of the visit

The purpose of my visit would be to expand on the development of research tools I have created as well as using these tools to test the roles of grammar and lexical information on morphophonological alternations. The research group in Ghent has great expertise in developing resources for psycholinguistic research (e.g. Brysbaert & New, 2009; Keuleers & Brysbaert, 2010; Keuleers et al. 2012) and has also contributed to research in morphophonology (e.g., Keuleers et al, 2007) I have recently developed a Brazilian Portuguese frequency corpus using film subtitles, following the footsteps of this group (e.g., Keuleers, Brysbaert & New, 2010). The corpus was used to test the role of the grammar (monosyllabicity, and initial syllable constraint) versus the lexicon (token frequency) on morphophonological alternations in the plural construction in Brazilian Portuguese (Becker et al., 2012).

The development of carefully tested research tools benefits the research community by raising the overall quality of experimental stimuli and analyses. The Ghent group is spearheading an open collaborative effort (the openlexicons project) that aims to reduce the time required to develop tools required for experimental linguistic research in languages for which these resources are not yet available (e.g., validated word frequencies, tests for lexical knowledge, pseudoword generators). For this, they require external collaborators' sound knowledge of computational techniques and expertise in linguistics. My educational background in engineering would allow me to contribute towards duplicate-removal technique and automation in general. My knowledge of Chinese would contribute to the

development of research tools for that language, such as corpus normalization, which requires native-speaker knowledge. Second, as there is not a comprehensive methodology for creating pseudowords for character-based languages such as Mandarin Chinese, and as there is a growing interest in psycholinguistic research in comparing East Asian languages with European ones, I believe that with the intimate knowledge this group has on pseudo words, a preliminary development should be possible even in short period of time. In addition, my working knowledge of Portuguese would allow us to develop resources for that language.

Turning to modelling possibilities, the collaboration would involve the application of memory-based learning principles that have been applied to Dutch plural formation (e.g. Keuleers et al, 2007, Keuleers & Daelemans, 2007), to our morphophonological alternation data, thereby bring a closer fit to the development loop between corpus construction and quantitative resources, theoretical models of psycholinguistic storage and retrieval, and experimental results from wug-tasks and lexical decision batteries.

2) Description of the work carried out during the visit

Three sub-projects were carried out during the visit which all falls under the openlexicons project:

1. Brazilian Portuguese Subtlex Project
2. Sound Symbolism
3. Pseudo-word construction for Mandarin Chinese

Brazilian Portuguese Subtlex Project

Tang (2012) documented how a Subtlex corpus can be enriched beyond tabulating the token frequency of words, e.g. Neighbourhood density, Lemmatisation, Part-of-Speech tagging, Grapheme to Phone conversion, and N-gram versions of the corpus. We investigated the availability of the tools needed for these post-processing and implemented them. Furthermore we implemented a Brazilian Portuguese module for creating pseudo words using Wuggy (Keuleers, E., & Brysbaert, M., 2010), the lexicon for such a module was based on the word list from Subtlex Brazilian Portuguese.

Lemmatisation and Part-of-Speech tagging: the TreeTagger for Portuguese by Pablo Gamallo was used, http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

Neighbourhood density : Orthographic Levenshtein distance 20 (OLD20) (Yarkoni et al., 2008) which is the average Levenshtein distance of the 20 closest neighbours was calculated, it was chosen because it has been suggested to be a better metric than Colheart's N in predicting performance in behavioural tasks.

Grapheme to Phone conversion: there is no readily available converter for Brazilian Portuguese, so a European Portuguese converter was used, with added hard-coded rules (in progress). http://www.co.it.pt/~labfala/g2p/ (Signal Processing Lab, Instituto de Telecomunicacoes)

N-gram: By creating a bigram word frequency norm, we could now search for potential compounds and collocation frequency.

Wuggy (Brazilian Portuguese module): Such a module requires only a syllabified word list (orthography) and a list of possible orthographical nuclei. Brazilian Portuguese syllabification was performed using Lingua-PT-Hyphenate Perl Module by Jose Alves de Castro.

Sound Symbolism in English

Sound symbolism has been a long debated topic. Many studies have approached this using cognates, phonetic properties of sounds and more. We approached this using a reconstruction approach with topic modelling.

The text corpus we used was a bigger version of SUBTLEX-US. To avoid potential artefacts in our analyses, the corpus were first lemmatised then morphemised. These artefacts are due to the fact that the inflected forms of a lemma will have similar semantic content as well as phonetic content, e.g. laugh-ing and laugh-ed. Part Of Speech tagging, and lemmatisation were performed using Stanford-Lemmatiser. To break the lemmas into morphemes, we used the morphological breakdown of polymorphemic words from CELEX [Baayen et al., 1995], and broke each lemma down into its smallest decomposition, e.g. the word unnecessarily would be broken down into three morphemes un, necessary, and ly. Stop word and stop morphemes were removed before applying topic modelling.

In analyses of semantics in machine learning, there exists two prevalent techniques – Latent Semantic Analysis (henceforth LSA) [Landauer and Dumais, 1997] and Latent dirichlet allocation (henceforth Topic Model) [Blei et al., 2003] In a comparison of the two techniques by Griffiths et al. [2007], the topic model performed better in predicting word association and a range of linguistic processing and memory tasks. For this reason, the Topic Model was chosen for extracting the semantic representation of words from our corpus. Finally, a topic model was performed on the lemmatised-morphemized corpus with 400 topics (in later tests, we tried 800, 1200 topics as well). As a first step, used only monomorphemic and monosyllabic words.

Three different distance/similarity metrics were explored on orthographical and phonemic forms. A segmental metric, Levenshtein distance [Levenshtein, 1966] on the orthography and phonemes. Since it is blind to featural differences: Manner, Place of Articulation and Voicing, we also employed two featural based metrics, ALINE (Local Alignment) [Kondrak, 2002, Huff, 2010] and Modified Value Difference Metric (henceforth MVDM) [Cost and Salzberg, 1993, Keuleers and Daelemans, 2007]

Using a leave-out-approach, we reconstructed the semantic vector for each word using only the semantic vectors of the remaining words, weighted by their corresponding phonetic similarity with the words that are being reconstructed. Furthermore, we applied different weighting schemes, 1) None (directly use the phonetic similarity values of between all the words and the reconstruction word), 2) A weighting decay function [Shepard et al., 1987] was applied to the phonetic similarity, 3) A hard-cut-off, we used only the phonetic similarity of the n-th closest neighbours [Luce and Pisoni, 1998, Yarkoni et al., 2008].

To evaluate our results, we examined a) if the reconstructed semantic space is correlated more with the original semantic space than a randomly reconstructed semantic space in terms of the relative semantic similarity between all the words b) the

reconstructability of widely-reported symbolically-motivated words, e.g. glow, gleam, crash, bash, tweet etc.

We also approached the reconstruction from the other direction, reconstructing sound from meaning, again using the leave-one-out approach.

Pseudo-word creation for Mandarin Chinese

Pseudo-words play a crucial part in lexical decision tasks. The quality of pseudo-words is vital for avoiding potential biases, in alphabetic languages, this has been very well controlled for, using bigram/trigram strings, calculating statistics such as neighbour density and more.

In Mandarin Chinese, the prevalent methods are to create pseudo-word and pseudo-characters. Currently, Chinese pseudo-words (which are predominantly dissyllabic) are generally created by doing character swapping of one or both characters, if the resultant combination of characters does not exist in a dictionary, then it is considered to be a pseudo-word. Similarly, Chinese pseudo-character are created by doing radical/grapheme swapping; since a lot of Chinese characters are composed of two graphemes, one or both graphemes would be swapped with another existing grapheme, and if the character does not exist in a dictionary, then it is considered to be a pseudo-character.

We argued that neither of the existing methods are holistic/principled ways of creating pseudo-word/characters in Chinese. Firstly, using existing characters in the swapping process has the problem that characters can also be words, so the fact that a particularly combination of the characters does not exist in a dictionary does not mean this combination cannot be a word. Secondly, while using existing graphemes in the swapping process is a principled way for creating pseudo-characters, it is far from perfect, as many characters contain more than two graphemes, and graphemes/characters can be embedded in other characters.

Our proposed approach was to break all the characters down into a tree-structure, using the CJK library, and to apply a similar algorithm to that of Wuggy.

3) Description of the main results obtained

Brazilian Portuguese Subtlex Project

Our creation of a very large subtitle corpus for Brazilian Portuguese, openly available and in a standardized format, will remain accessible as a potentially valuable resource for phonological & psycholinguistic research and for a number of adjacent fields. We demonstrated the richness of subtlex corpora and their potential for research beyond token frequency, ranging from lexical neighbourhood to pseudo-word creation.

Different versions of the corpus (with different filters) with an interactive interface are available at http://crr.ugent.be/subtlex-pt-br/

For more specific corpora:
Unigram:
http://zipf.ugent.be/open-lexicons/interfaces/pb-subtitles-unigram/
Bigram:
http://zipf.ugent.be/open-lexicons/interfaces/br-pt-bigrams/
Lemmatised + POS-Tagged:

http://zipf.ugent.be/open-lexicons/interfaces/br-pt-lemmas/

Sound Symbolism in English

The contribution of sound symbolism in the English lexicon is small, but the link between sound and meaning is not arbitrary, but shown to be above chance. There's a clear locality effect of neighbours. Our results confirmed native speakers' intuitions of sound symbolism with the skewed distribution of allegedly symbolically-motivated words in terms of reconstructability.

Metrics applied on orthography outperformed those applied on phonemes in many models. This is perhaps due to the fact that the English spelling does not generally reflect the sound changes in the pronunciation, consider, e.g. ni**gh**t–lau**gh**, **g**naw, lam**b** etc.

We consistently found that by applying weighting schemes to the reconstruction, the correlation value is higher than with no weighting schemes at all. This suggests that there's a locality effect of neighbours, that is, closer neighbours contribute more to sound symbolism than distant neighbours.

By comparing the semantic similarity of all the words in the reconstructed space and those in the original space, we found that the reconstructed space is consistently more correlated than a randomly reconstructed space. This holds true across all the conditions that we tested (phonetic similarity metrics, weighting schemes and topic sizes). Furthermore, we extracted a list of widely-reported symbolically motivated words in English. It was found that these words are indeed more reconstructable.

Together, these findings provide ample evidence that the link between meaning and sound is not arbitrary.

Pseudo-word creation for Mandarin Chinese

We encountered two technical problems which are solvable but is beyond the scope and time of this short visit. Firstly existing databases which contain the breakdown of the characters as tree structures are unspecified, and the mapping between the characters and their tree structures is a many-many relationship. More than one character can have the same tree structure and contents, and more than one tree structure can capture a character.

Secondly, there is no existing system to print characters using tree structure as the input. We explored a few potential systems but they are either out-dated, discontinued, or using a different input system.

4) Future collaboration with host institution (if applicable)

- Further work on Sound Symbolism but with different languages
- Continuing to contribute to the Openlexicons project

5) Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant)*

Yet to be discussed

6) Other comments (if any)

I would like to thank the ESF for funding my visit. I have learnt more than could have done by myself. I have brought back skills and techniques I have learnt from University of Gent to University College London. There are plenty of potential ideas for collaboration between the two labs. It was a pleasure working with the CRR group in Gent, particularly with Dr. Emmanuel Keuleers and Pawel Mandera.

Reference:

M. Becker, L.E. Clemens, and A. Nevins. A richer model is not always more accurate: the case of French and Portuguese plurals. Lingbuzz, 2012. URL http://ling.auf.net/lingBuzz/001336.

M. Brysbaert and B. New. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior Research Methods, 41(4): 977–990, Nov 2009.

H.R. Baayen, R. Piepenbrock, and L. Gulikers. The CELEX Lexical Database. Release 2 (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, 1995.

D.M Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

T.L. Griffiths, M. Steyvers, and J.B. Tenenbaum. Topics in semantic representation. Psychological review, 114(2):211, 2007.

L. Hinton, J. Nichols, and J.J. Ohala. Sound symbolism. Cambridge University Press, 2006.

Tang, K. (2012) A 61 Million Word Corpus of Brazilian Portuguese Film Subtitles as a Resource for Linguistic Research. UCL Working Papers in Linguistics 24.

E. Keuleers and M. Brysbaert. Wuggy: A multilingual pseudoword generator. Behavior Research Methods, 42(3):627–633, 2010.

E. Keuleers and W. Daelemans. Memory-based learning models of inflectional morphology: A methodological case-study. Lingue e linguaggio, 6(2):151–174, 2007.

E. Keuleers, M. Brysbaert, and B. New. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. Behavior Research Methods, 42(3):643–650, Aug 2010.

E. Keuleers, D. Sandra, W. Daelemans, S. Gillis, G. Durieux, & E. Martens (2007). Dutch plural inflection: The exception that proves the analogy. Cognitive Psychology, 54(4), 283–318.

V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet Physics Doklady, volume 10, pages 707–710, 1966.

P.A. Luce and D.B. Pisoni. Recognizing spoken words: The neighborhood activation model. Ear and hearing, 19(1):1, 1998.

R.N Shepard et al. Toward a universal law of generalization for psychological science. Science, 237(4820):1317–1323, 1987.

T. Yarkoni, D. Balota, and M. Yap. Moving beyond Colthearts N: A new measure of orthographic similarity. Psychonomic Bulletin & Review, 15(5):971–979, 2008.