Natural Language Processing and Cognitive Science

Proceedings 2018

Bernadette Sharp, Wiesław Lubaszewski
and Florence Sedes (eds.)

Bernadette Sharp, Wiesław Lubaszewski
& Florence Sedes (eds.)

Natural Language Processing and Cognitive Science

Proceedings 2018

Jagiellonian Library
Kraków 2018

ISBN 978-83-949716-7-0
eISBN 978-83-949716-8-7

Editors

Bernadette Sharp, Staffordshire University, U.K.
Wiesław Lubaszewski, Jagiellonian University, Poland
Florence Sedes, Paul Sabatier University, France

Typesetting and pake makeup:
Martyna Fatel

Cover design:
Katarzyna Szczepaniec

Biblioteka Jagiellońska
al. Mickiewicza 22
30-001 Kraków
tel. 12 633 09 03
www.bj.uj.edu.pl

Contents

13th International Workshop on Natural Language Processing and Cognitive Science NLPCS 2018

11-12 September 2018
Jagiellonian University, Krakow, Poland

Research into natural language processing has never been as exciting as in the last few years reflected by the significant increase in the amount of conferences, workshops and events dedicated to this field. We can see some overlap in the topics of natural language conferences with those covered by Artificial Intelligence events. NLPCS, which was launched in 2004, have provided a strong platform acknowledging the importance of interdisciplinary approaches; the aim is to bring together computer scientists, cognitive and linguistic researchers to improve our understanding of the human language system.

There is an increasing interest in research related to the human brain and human language, and their findings and advances will contribute to advance our study of natural language processing. The aim of this 13th international workshop of Natural Language and Cognitive Science (NLPCS 2018) is to explore the special relationship between natural language processing, cognitive science and cognitive neuroscience. Cognitive science is the study of mind and intelligence. Cognitive neuroscience is concerned with the study of the biological processes and aspects that underlie cognition focusing in particular on the neural connections in the brain which are involved in mental processes.

This year we had the pleasure to welcome Professor Roelien Bastiaanse (University of Groningen, The Netherlands), who gave her talk on "How language impairments can help us to understand how language is processed in the brain: time reference in agrammatic aphasia".

NLPCS workshops has produced the following publications in addition the proceedings:

- Sharp, B., Sedes, F. & Lubaszewski, W. (Eds.) *Cognitive Approach to Natural Language Processing,* ISTE Elsevier, 2017
- Gala, N., Rapp, R. & Bel-Enguix, Gemma (Eds.) *Language Production, Cognition, and the Lexicon*, Springer, 2014
- Neustein, A. & Markowitz, J.A. (Eds.) *Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems*, Springer Verlag, Heidelberg/New York, 2013
- Special issue of the *International Journal of Speech Technology*, vol. 12, 2/3, September 2009
- Special issue of the *International Journal of Speech Technology*, vol. 11, issue 3/4, December 2008

I would like to thank the programme committee for their help and support and specially the organising committee and authors who have contributed to make this event possible.

NLPCS 2018 co-chairs:

Bernadette Sharp (Staffordshire University, United Kingdom, b.sharp@staffs.ac.uk)
Wieslaw Lubaszewski (Jagiellonian University, Poland, lubaszew@agh.edu.pl)
Florence Sedes (Paul Sabatier University, France, florence.sedes@irit.fr)

NLPCS Programme Committee:

- Aretoulaki, M. (Dialog Connection, UK),
- Barnden, J.A. (University of Birmingham, UK),
- Cristea, D. (University of Iasi, Romania),
- Finley G. (University of Minnesota, USA),
- Gatkowska, I. (Jagiellonian University, Poland),
- Korzycki, M. (Industry, Poland),
- Lee M. (University of Birmingham, UK),
- Noriega L. (Manchester Business School, UK),
- Oakes M. (Wolverhampton University, UK),
- Rayson, P. (Lancaster University, UK),
- Roche, C. (Université de Savoie, France),
- Schwab, D. (Universite de Grenoble Alpes, France),
- Schwitter, R. (Macquarie University, Australia),
- Suendermann-Oeft, D. (Dialog, Multimodal, and Speech Research Center).

What Language Impairments Can Tell Us about the Organization of Language in the Brain

Roelien Bastiaanse[1,2]

[1] Center for Language and Cognition Groningen (CLCG), University of Groningen, PO Box 716, 9717 GB Groningen, The Netherlands, y.r.m.bastiaanse@rug.nl
[2] Center for Language and Brain, National Research University Higher School of Economics, Russian Federation

Abstract Agrammatic aphasia is characterized by non-fluent speech in which free and bound grammatical morphemes are frequently omitted or substituted. However, not all grammatical morphemes are equally vulnerable. The present paper focuses on grammatical morphemes that are used for time reference. For this, some languages used verb morphology, others use grammatical tone or free morphemes, so-called 'aspectual adverbs'. Data from nine different languages will be presented and it will be shown that reference to the past is selectively impaired in agrammatic speakers of languages that use verb inflections for time reference. However, in languages that use aspectual adverbs, this selective deficit does not occur. These results will be discussed in relation to the linguistic concept of Discourse Linking.

1. Introduction

Aphasia is a language disorder due to brain damage; this brain damage can be cause by a stroke, traumatic injury, a brain tumor or, rarely, infections. The most common cause in the Western world is a stroke. The individual characteristics and of aphasia are dependent on the site and size of the lesion, that is most frequently in the left hemisphere. The linguistic levels – grammar, lexical-semantics and phonology – can be independently affected. The most commonly studied form of aphasia is 'agrammatism' or Broca's aphasia, a grammatical disorder that is classically described as non-fluent speech consisting of mainly content words (nouns, verbs, adjectives) with omissions and / or substitutions of free and bound grammatical morphemes (e.g., Goodglass & Kaplan, 1972). Language comprehension usually remains intact, although grammatically complex sentences (semantically reversible passives, object clefts and object relatives) are often poorly understood (see e.g., Grodzinsky, 2000; Yarbay Duman & Bastiaanse, 2009). An example of agrammatic speech is given in (1). [The words of the interviewer are between square brackets; …. pauses].

(1) Amsterdam … and eh … beautiful … eh … I … nice … walk [Okay. Where?] Where? Eh … Amsterdam [Are you walking around the city?] No bike or no eh … eh … car eh … shopping … and eh … eh call and eh … first eh … eh … cup of coffee … eh … Mary and eh … talking a bit.

The classic definition of Broca's aphasia is not entirely correct. It has repeatedly been shown that in agrammatic aphasia, production of verbs quite selectively affected, not only in spontaneous speech (Thompson et al., 1994; Bastiaanse & Jonkers, 1998), but also on an action naming task (Luzzatti et al., 2002; Jonkers & Bastiaanse, 2007). The question is: why is the production of verbs difficult? Several suggestions have been made: their argument structure (Kim & Thompson, 2002); the position of finite verbs in the syntactic tree (Friedmann & Grodzinsky, 1997); the fact that finite verbs, in some languages like Dutch and German, are in the derived position (Bastiaanse & Van Zonneveld, 1998; 2004); or the combination of hampered verb retrieval and verb inflection (Bastiaanse & Jonkers, 1998). Despite the discussion on the underlying cause of the problems with verbs, there is agreement that finite verbs forms are most severely affected.

In 2008, Bastiaanse showed that not all finite verb forms are equally affected, nor are all non-finite forms (including infinities and participles) equally spared. She found that the production of verb forms referring to the past (past tense; participles) were more impaired than verb forms referring to the present (present tense; infinitives). This was the beginning of a large and still ongoing crosslinguistic project to the problems with time reference in agrammatic aphasia, which includes data of more than 15 languages.

This project aimed to answer several questions:
(1) is only reference to the past impaired or to the future as well?
(2) can we find this selective impairment across languages that use verb inflection
and what is the effect of different morphological ways to refer to the past?
a. simple verb forms
b. periphrastic verb forms
c. grammatical tone
(3) is time reference affected in languages that do not have verb inflection, like Chinese,
Standard Indonesian and Thai, but use aspectual adverbs?

The first question we addressed in a study on Turkish agrammatic speakers. Turkish is, an agglutinative language in which time reference, in most time frames, is done by inflectional tense. Yarbay Duman et al. (2009) compared production of reference to the past and future and found that Turkish agrammatic speakers were significantly more impaired in the past condition than in the future condition. Combined with the Dutch data of Bastiaanse (2008), who compared past and present tenses, this suggested that reference to the past was selectively impaired. Based on these data, the PAst DIscourse LInking Hypothesis (PADILIH; Bastiaanse et al., 2011) was launched. This hypothesis was based on two theories. First there was the linguistic theory of Zagona (2003), based on ideas Reichenbach (1947) and Partee (1979), that in present tense the time of speaking and the time of the event coincide, whereas in the past tense the time of speaking and the time of event differ. In order to connect the time of speaking to the time of the event in the past, discourse linking is required. Zagona (2003) considers time reference as a dual concept: there is [+past] for which discourse linking is needed and [-past] that includes present and future and which requires no discourse linking. The second theory comes from Avrutin (2006) who argues, that discourse linked element, such as pronouns (as opposed to reflexives), are difficult for agrammatic speakers. In Bastiaanse (2011) these two theories have been adjusted and combined to formulate the PADILIH: reference to the past through verb morphology is difficult for agrammatic speakers. Notice that this is slightly different from what Zagona (2003; 2013) argues: she only included past tense in her theory, whereas the PADILIH considers all verb forms that refer to the past, including, for example, the present perfect in Dutch and German (*hij heeft een boek gelezen:* lit. 'he has read a book') to be difficult for agrammatic speakers, even though the finite verb (*heeft*: 'has) is in present tense.

The PADILIH has been tested in a number of languages. For the current paper, those data that can answer the questions mentioned above are described, that is, we evaluate whether the PADILIH correct predicts the agrammatic problems with reference to the past in language that use different verb morphology (tensed verbs, periphrastic verb forms, grammatical tone) or different grammatical morphology (aspectual adverbs) to refer to the past.

2. Methods

2.1 Participants

For all languages, first a number of non-brain-damaged speakers, age and education matched with the agrammatic speakers, were tested, to see whether the items elicited the intended responses. That was the case for all languages: this population scored at ceiling.

The aphasic participants were selected on the basis of agrammatic speech. Notice that in most languages that are included, no standard aphasia test is available to determine aphasia type. If there was, only individuals with Broca's aphasia were included, provided that they were able to talk. In all other languages the criteria for inclusion were: slow and effortful speech, characterized by a lack of function words and problems with grammatical morphology (so-called 'telegraphic speech'); auditory comprehension of words, which was measured by the (for some languages culturally adapted version of) subtest of the BDAE (Goodglass & Kaplan, 1972), should be intact.

2.2 Materials

PADILIH has been tested for typologically different languages. For most languages, the same test, *Test for Assessing Reference of Time (TART;* Bastiaanse, Jonkers, & Thompson, 2008) has been used. For some languages, the TART had to be adapted, which will be mentioned. We used a sentence completion test, for which the participant was prompted with the intended structure. Two photos of contrasting actions (e.g. eating - drinking; pushing - pulling) were shown and the verb belonging to the action was printed above the picture. Within one item, both actions were completed, ongoing or commencing (see Figure 1). The experimenter made a sentence describing the first picture and started to describe the other picture; the agrammatic speakers was asked to finish the sentence. He had to produce the verb (complex) and the object. For example, for the items in Figure 1, the following sentence (2) was given:

(2)Experimenter: 'This picture is about peeling and this picture is about eating. Here you can say the man just peeled the apple and here you can say the man just ….'

Participant: '…. ate the apple'

The white models in the pictures were replaced by black and Asian ones for the African and Chinese version of the TART and some additional cultural changes were made (e.g. in the item mentioned in (2), the apple was replaced by an orange in the African version).

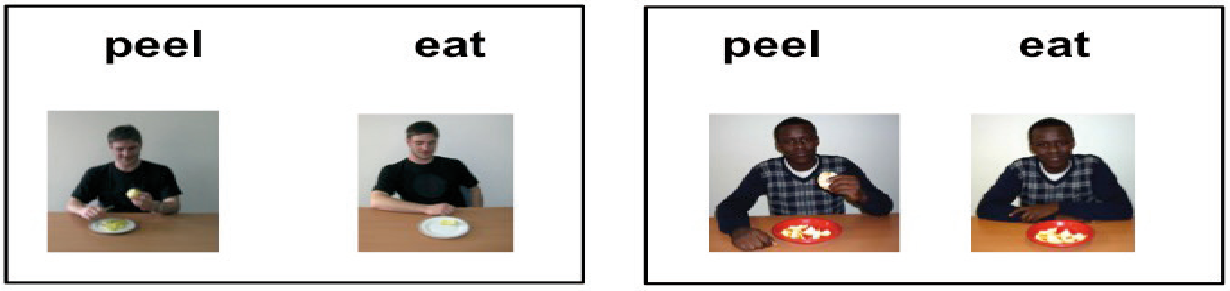


Fig. 1. Item in the Western and African version of the TART belonging to the sentence in (1).

2.3 Procedure and Scoring

The agrammatic speakers were tested individually in a quiet room. After two examples in each time frame in which feedback was given ('yes, this is correct'; 'no, this is not entirely correct, the answer should be ….'), the test started. No feedback or correction was provided anymore. The entire session was recorded and the answers were later transcribed and scored both quantitatively and qualitatively. Here an overview of the quantitative results is given; the results of the qualitative analysis can be found in the original papers.

2.4 Statistical analysis

The group sizes differ per language and statistical testing was adjusted accordingly. The group sizes are mentioned in the captions of the figures. The differences that are mentioned below are all statistically significant ($p<0.05$). If no difference is found, this means that $p>0.05$.

3. Results

3.1 Time reference through verb morphology

English (Bastiaanse et al., 2011) For English three conditions were tested: simple past (ate), present continuous (is eating) and future (will eat). The results are given in Figure 2. These results support the PADILIH, but notice that these results might be interpreted differently: the data also show that simple verb forms ('ate') are more difficult than periphrastic verb forms ('is eating'; 'will eat'). There are two ways to exclude this latter explanation: doing the same experiment in a language in which all three verb form are simple verbs (like Turkish) or comparing two time frames that can be referred to by both a simple and a periphrastic verb form within one language (like Dutch). We did both.

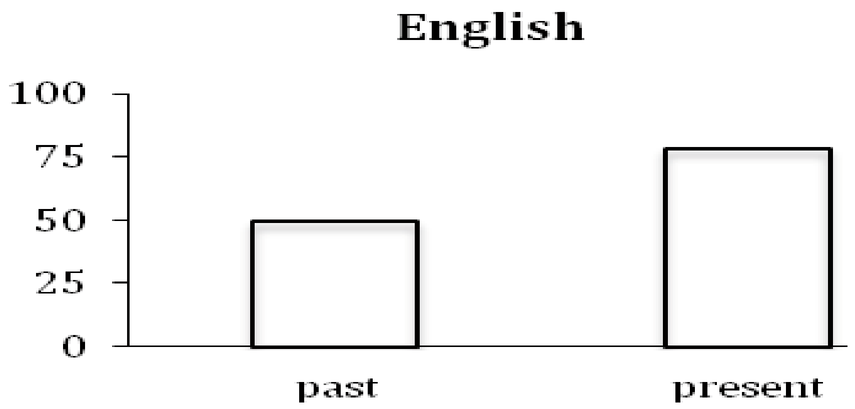


Fig. 2. The percentages correct for English (n=12); difference between past and present: p<0.05.

Turkish (Bastiaanse et al., 2011; Arslan et al., 2014) Turkish is an agglutinative language in which most verb forms are single words. In Bastiaanse et al. (2011), we report on the use of the TART for testing past, present and future time reference and, as is shown in Figure 3a, the PADILIH was supported: past time references was selectively impaired.

However, this is not the whole story. Turkish has different verb forms to refer to the past dependent on whether or not the speaker witnessed the event. If the speaker did, the so-called 'direct evidential' verb form [verb stem = DI] is used, but if the speaker heard about the event or if he made inferences about the event, the so-called 'reportative' and 'inferential' verb form [verb stem = mIş] is the correct form. Arslan et al. (2014) argue that for the direct evidential the event time and the evaluation time are in both the past, and hence, discourse linking is required. For both the reportative and inferential verb form, the event time is in the past, but the evaluation time coincides with the time of speaking Thus, PADILIH predicts that the direct evidential will be more impaired than the other two forms. A new test was designed with video clips resembling the TART (see original paper). As shown in Figure 3b, the direct evidentials, that had been witnessed by the participants, were more difficult than the reportative and inferential verb form.

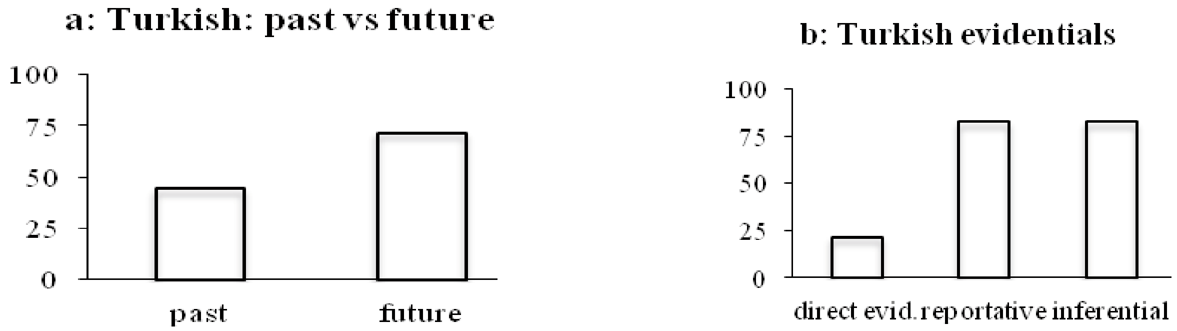


Fig. 3. The percentages correct for Turkish: (a) past compared to future (n=8; Bastiaanse et al., 2011; p<0.05). (b) direct evidentials compared to reportative and inferential (n=7; Arslan et al., 2014; difference between direct evidentials and other two forms: p<0.05).

Dutch (Bos & Bastiaanse, 2014) In order to test the influence of the verb form, that is, whether reference to the past is made by a single or a periphrastic verb form, the Dutch TART was used. In Dutch, there are two ways to refer to the past: the past imperfective, which is a single verb in past tense, and the present perfect which is a periphrastic verb form with the auxiliary in present tense and the lexical verb as an participle (see 3a-b).

(3a) de manat een sinaasappel
the man ate an orange
(3b) de man heeft een sinaasappel gegeten
the man has an orange eaten

The sentences mean the same and are equally frequent. The results (see Figure 4) show that the periphrastic and simple past both more difficult that the present. This demonstrates that it is not the past tense, that is affected, but rather verb forms that refer to the past, even if the finite verb is in the present tense.

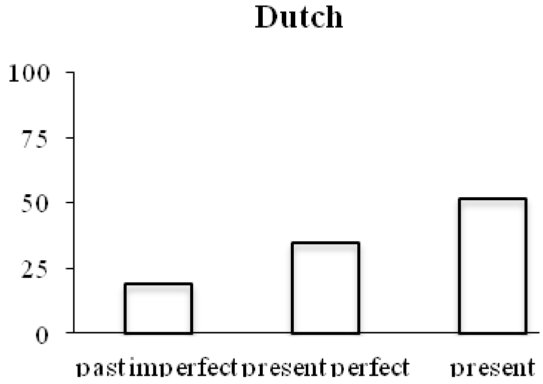


Fig. 4. The percentages correct for Dutch (n=14): past and present perfect are both worse than present ($p<0.05$); there is no difference between the two forms referring to the past (present perfect and past imperfect: $p>0.05$).

3.2 Time reference through grammatical tone

Akan (Tsiwah & Bastiaanse, in press) Akan is a tone language spoken in Ghana and Ivory Coast. Interestingly, Akan has grammatical tone that is (among others) used for distinguishing time frames. In (4a-b) this is illustrated: ` indicates low tone and ´ indicates high tone.

(4a) Peter gyìná hɔ
Peter stand-HAB there.
'Peter stands there'
(4b) Peter gyìnàà hɔ
Peter stand-PAST there.
'Peter stood there'

The past (4a) and present habitual (4b) are not distinguished by a grammatical morpheme, but rather by tone and duration of the final vowel. To find out whether it is time reference to the past through bound and/or free morphemes that makes reference to the past difficult or that grammatical tone, agrammatic Akan speakers were tested. As can be seen in Figure 5, grammatical tone is as vulnerable as grammatical morphology and the pattern is the same as in other languages.

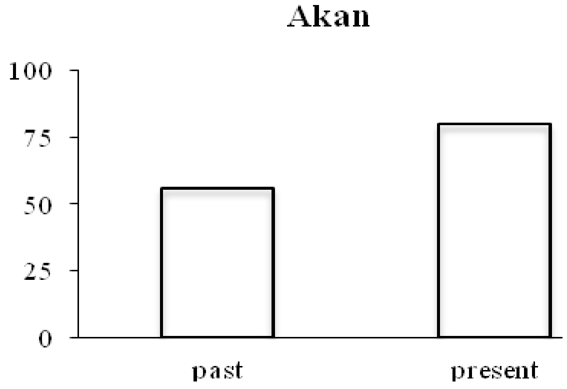


Fig. 5. The percentages correct for Akan (n=9; $p<0.05$).

This shows that not only verb forms referring to the past which are marked by free or bound inflectional morphology are affected. Rather, it is reference to the past through inflectional or morphology and tone that is vulnerable in agrammatic aphasia. This is all in line withthe PADILIH, which predicts that reference to the past is impaired because it is discourse linked.

3.3 Time Reference through Aspectual Adverbs

So far, we have shown that the PADILIH correctly predicts that reference to the past by verb inflection, be it through free and bound morphemes or tone, is selectively affected in agrammatic aphasia. However, there are languages that do not mark their verbs for time reference, but use so-called 'aspectual adverbs', such as Chinese (5a-b). These adverbs are only used when it is not clear from the context whether an event has been completed, is ongoing or will happen in the future.

(5a) zhe ge ren   du   le   yi fong sin
the man   read   [perfect]   a letter
'the man read the letter'
(5b) zhe ge ren   zai   du   yi fong sin
the man   [dur]   read   a letter
'the man is reading a letter'

Chinese (Bastiaanse et al., 2011) The first language with aspectual adverbs we tested was Chinese. The instructions were: 'for this picture you can say the man just [read + perfect marker] the letter and for this picture you can say the man just ….' and the agrammatic speaker was supposed to answer '… [write + perfect marker] the letter'. The Chinese agrammatic speakers were remarkably poor in both conditions and no difference between past and present time reference was found (see Figure 6). On hindsight, this is not surprising. Unlike verb morphology, the aspectual adverbs are not obligatory, which was clear from the data: the agrammatic speakers (but not the non-brain-damaged control speakers) omitted the aspectual adverbs, which resulted in grammatical sentences that adequately referred to the right time frame. This showed that the TART in this form was not the right instrument to test time reference in Chinese.

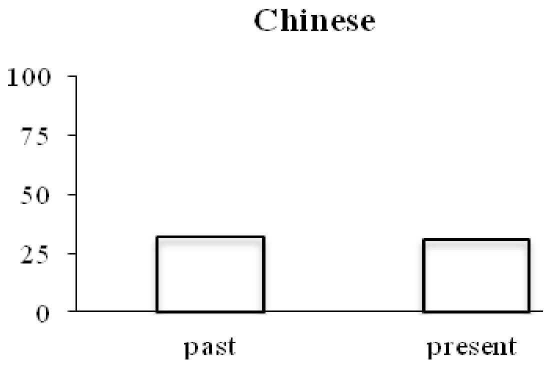


Fig. 6. The percentages correct for Chinese (n=11; p>0.05).

Standard Indonesian (Anjarningsih, 2012) Standard Indonesian also uses aspectual adverbs for time reference (see 6a-b).

(6a) Dia   sudah   menyapu   lantai.
She   perfect sweep   floor
'She swept the floor'
(6b) Dia   sedang menyapu   lantai.
She   durative sweep   floor
'She is sweeping the floor'

For Indonesian the TART was adjusted. Instead of contrasting the verbs on both pictures, time frames were contrasted, so that use of aspectual adverbs was required: 'for this picture you can say she [perfect marker + sweep] the floor and for this picture you can say she ….' and the agrammatic speaker was supposed to answer '… [durative marker + sweep] the floor.' The results are given in Figure 7. As can be seen, the Standard Indonesian agrammatic speakers performed similarly for reference to the past and to the present. The question is why is reference to the past not different from reference to the present. The reason may be in the optional nature of the use of aspectual adverbs: they are only used when the time frame is not clear from the context. Basically, this means that they are used when the time of speaking needs to be linked to the time of the event, be it past, present or future. It is, hence, suggested that time reference through optional aspectual adverbs is discourse linked for all time frames.

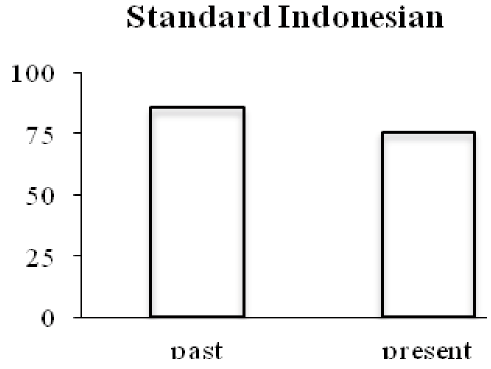


Fig. 7. The percentages correct for Standard Indonesian (n=5); no difference between past and present (p>0.05).

The PADILIH was adapted to these findings and formulated as: time reference through grammatical morphology for which discourse linking is needed, be it verb inflection, grammatical tone or aspectual adverbs, is impaired in agrammatic aphasia. If this is true, the same results should be found in other languages with aspectual adverbs. One of them is Thai.

Thai (Siriboonpipattana, in progress) Thai uses aspectual adverbs, just like the two previously described languages (see 7a-b).

(7a) tɔɔn-niĭ   khǎw   dʉʉm   nom   yuù
now   he   drink   milk   PRES
'Now, he is drinking milk'
(7b) mʉǎ-waan   khǎw   dʉʉm   nom   lɛ́ɛw
yesterday   he   drink   milk   PAST
'Yesterday, he drank) milk'

Again the TART was adapted: now three pictures were presented to the participant, e.g., one of a man drinking milk, one of a man about to drink milk and one of a man with an empty glass that obviously contained milk. One of these pictures had a red frame and the agrammatic speaker had to describe that picture in one sentence so that the experimenter, who could not see the picture, was able to select the correct picture. This can only be done when the correct time reference is made. Although the number of participants is still low, the results so far are as predicted (see Figure 8). Statistical analysis is too early, but it looks as though both tested time frames are equally affected and, thus, support the adjusted PADILIH.

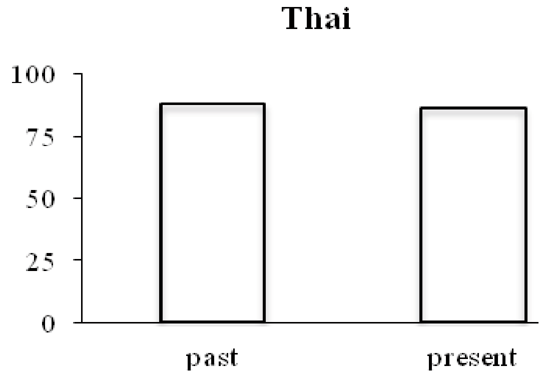


Fig. 8. The percentages correct for Thai (n=3); no difference between past and present.

3.4 Time Reference in Bilingual Agrammatic Speakers

When a person who speaks two or more languages develops aphasia, usually all languages are equally affected. Testing balanced bilingual aphasic speakers is the perfect way to make cross-linguistic comparisons. However, balanced bilingual speakers are usually hard to find and agrammatic balanced bilingual speakers are even more rare. However, in many African countries, people speak multiple languages. For our study we focused on bilinguals in Kenya. During the first years of their lives, children usually learn one of the many indigenous languages. When they go to school the learn English and Swahili, which they use equally on a daily basis for the rest of their lives.

Swahili-English Bilingual Aphasia (Abuom & Bastiaanse, 2013)

Swahili is an agglutinative language and the verbs are morphologically complex: pronouns, tense, negation, voice are all included in the finite verb. The tense morpheme is one of the prefixes, as shown in (8a-b).

(8a)	mwanamme	[a	# **na** #	andika]	barua
	man	[subject prefix	present	write]	letter
	'the man is writing a letter'				
(8b)	mwanamme	[a	# **li** #	andika]	barua
	man	[subject prefix	past	write	letter
	'the man wrote a letter'				

The participants were tested with the African version of the original TART in both Swahili and English. The results are given in Figure 9.

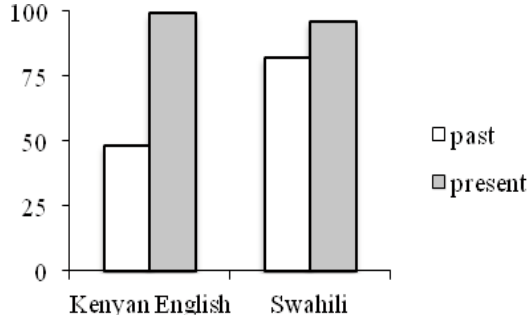


Fig. 9. The percentages correct for the bilingual speakers (n=13). Past is more difficult than present in both languages ($p<0.05$); English past is more difficult than Swahili past ($p<0.05$).

As can be seen in this figure, reference to the past is selectively impaired in both languages, as predicted by the PADILIH. Interestingly, reference to the past is more impaired in English than in Swahili, despite the complex verb inflection paradigm in the latter.

4. Discussion

This crosslinguistic project to time reference has shown that reference to the past through verb morphology is selectively impaired in agrammatic aphasia. The explanation for this is that for this kind of time reference discourse linking is required, which is notoriously difficult for agrammatic speakers (Avrutin, 2000). However, this only holds for languages that use obligatory verb morphology for time reference, be it inflectional morphemes or grammatical tone (like Akan); the selective impairment for reference to the past is not observable in languages in which optional aspectual adverbs are used. Since these aspectual adverbs are only used when the time frame is not clear, it is suggested that discourse linking is always required when the time frame is not clear from the discourse itself: that is why there is no difference between the use of aspectual adverbs referring to the past and the present in Chinese, Standard Indonesian and Thai. In Turkish, two forms can be used to refer to the past. Only the form that requires personal observation is supposed to be discourse linked and it is this form that is selectively impaired.

Not only grammatical morphology (verb inflection; grammatical tone; aspectual adverbs) can be used to refer to the past: lexical adverbs and adverbial phrases, such as *just*, *yesterday, last year, previously,* can also be used. Unfortunately, in languages that use verb inflection, these lexical adverbs cannot be torn apart from past tense. This can be done in languages that use aspectual adverbs, because in this case, a lexical adverb is enough to make the link to a time frame; aspectual adverbs are then no longer needed. We tested the production of temporal lexical adverbs in Standard Indonesian agrammatic speakers. Although the group of agrammatic speakers was quite small, the use of lexical adverbs seemed to be impaired as much as the use of aspectual adverbs (Anjarningsih, 2012). So far, we did not have the chance to study this more in depth.

The question is: what does this tell us about language organization in the brain? First of all, the data show how important linguistics is for studying language impairments due to brain damage. It is only recently that we begin to understand the underlying impairments in aphasia by applying our linguistic knowledge to language processing. For example, the ideas of Zagona (2003; 2013), combined with the theories of Reichenbach (1947) and Partee (1973) enabled us to understand why reference to the past may be impaired in agrammatic aphasia, a phenomenon that was only detected by accident in Bastiaanse (2008). The extra / different activation that is apparently needed to refer to the past has also been found in healthy speakers with ERP Dragoy et al., 2012; Bos et al., 2013); Eye Tracking (Bos et al., 2014a) and fMRI (Bos et al., 2014b).

In conclusion, these data show how important it is to study typologically different languages to test neurolinguistic theories. Studying aphasia is a suitable means to do this, even though it is always hard to find the right aphasia individuals. This is shown by the relatively small groups, although overall we tested more than 100 agrammatic speakers of more than 15 languages and the data all point in the same direction: time reference for which discourse linking is required is selectively impaired in agrammatic aphasia.

Acknowledgements

This project could not have been done without all my students and colleagues speaking so many different languages. They helped with the translations and adaptations the TART, recruited and tested the non-brain-damaged and aphasic individuals and designed the additional experiments for ERP, Eye Tracking and fMRI studies on time reference. I am also grateful to Katrina Gaffney for her comments on an earlier version of this paper.

The author is partially supported by the Center for Language and Brain NRU Higher School of Economics, RF Government grant, ag. № 14.641.31.0004.

References

Abuom, T.O. & Bastiaanse, R. (2013). Production and comprehension of reference of time in Swahili-English bilingual agrammatic speakers. *Aphasiology*, *27*, 921-937.

Anjarningsih, H.Y. (2012). *Time Reference in Standard Indonesian Agrammatic Aphasia.* Grodil, Groningen.

Arslan, S., Aksu-Koc, A., Mavis, I. & Bastiaanse, R. (2014). Finite verb inflections for evidential categories and source identification in Turkish agrammatic Broca's aphasia. *Journal of Pragmatics*, *70*, 165-181.

Avrutin, S. (2006). Weak syntax. K. Amunts, Y. Grodzinsky (eds.). *Broca's Region*, 49-62. Oxford Press, New York.

Bastiaanse, R. (2008). Production of verbs in base position by Dutch agrammatic speakers: Inflection versus finiteness. *Journal of Neurolinguistics*, *21*, 104-119.

Bastiaanse, R., Bamyaci, E., Hsu, C.-J., Lee, J., Yarbay Duman, T. & Thompson, C.K. (2011). Time reference in agrammatic aphasia: A cross-linguistic study. *Journal of Neurolinguistics*, *24*, 652-673.

Bastiaanse, R. & Jonkers, R. (1998). Verb retrieval in action naming and spontaneous speech in agrammatic and anomic aphasia. *Aphasiology*, *12*, 99-117.

Bastiaanse, R., Jonkers, R. & Thompson, C.K. (2008). *Test for Assessing Reference of Time (TART)*. University of Groningen, Groningen.

Bastiaanse, R. & Zonneveld, R. van (1998). On the relation between verb inflection and verb position in Dutch agrammatic aphasics. *Brain and Language*, *64*, 165-181.

Bastiaanse, R. & Zonneveld, R. van (2004). Broca's aphasia, verbs and the mental lexicon. *Brain and Language*, *90,* 198-202.

Bos, L.S. & Bastiaanse, R. (2014). Time reference decoupled from tense in agrammatic and fluent aphasia. *Aphasiology*, *28*, 533-553.

Bos, L.S., Dragoy, O., Stowe, L.A. & Bastiaanse, R. (2013). Time reference teased apart from tense: Thinking beyond the present. *Journal of Neurolinguistics*, *26*, 283-297.

Bos, L.S., Hanne, S, Wartenburger, I. & Bastiaanse, R. (2014a). Losing track of time? Processing of time reference inflection in agrammatic and healthy speakers of German. *Neuropsychologia*, *65*, 180-190.

Bos, L.S., Ries, J., Wartenburger, I., & Bastiaanse, R. (2014b). The neural correlates of reference to the past. *Frontiers of Psychology*, *5.*

Dragoy, O., Stowe, L.S., Bos, L. & Bastiaanse, R. (2012). From time to time: Processing time reference violations in Dutch. *Journal of Memory and Language*, *66*, 307-325.

Friedmann, N., & Grodzinsky, Y. (1997). Tense and agreement in agrammatic production: pruning the syntactic tree. *Brain and Language*, *56*, 397-425.

Jonkers, R. & Bastiaanse, R. (2007). Action naming in anomic speakers: Effects of instrumentality and name relation. *Brain and Language*, *102*, 262-272.

Goodglass, H. & Kaplan, E. (1972). *Boston Diagnostic Aphasia Examination*. Lea & Febiger, Philadelphia.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*, *23*, 1-21.

Kim, M., & Thompson, C.K. (2000). Patterns of comprehension and production of nouns and verbs in agrammatism: Implications for lexical organisation. *Brain and Language*, *74,* 1-25.

Luzzatti, C., Zonca, G., Pistarini, C., Contardi, A., & Pinna, G.D. (2002) Verb–noun double dissociation in aphasic lexical impairments: The role of word frequency and imageability. *Brain and Language*, *81*, 432-444.

Partee, B.H. (1973). Some structural analogies between tenses and pronouns in English. *The Journal of Philosophy*, *70*, 601-609.

Reichenbach, H. (1947). *Elements of Symbolic Logic.* 1st Free Press paperback, 1966 ed. Collier-Macmillan, London.

Thompson, C.K., Shapiro, L.P., Li, L., & Schendel, L. (1994). Analysis of verbs and verb argument structure: A method for quantification of agrammatic language production. *Clinical Aphasiology*, *23*, 121-140.

Tsiwah, F., Martínez Ferreiro, S., & Bastiaanse, R. (2018). Processing of time reference in a grammatical tone language speakers with agrammatic aphasia. *Presentation at the Science of Aphasia Conference XIX*, Venice, September 19-22.

Yarbay Duman, T. & Bastiaanse, R. (2009). Time reference through verb inflection in Turkish agrammatic aphasia. *Brain and Language*, *108*, 30-39.

Zagona, K. (2003). Tense and anaphora: Is there a tense-specific theory of coreference. In Barrs, A. (ed.) *Anaphora: A Reference Guide*, 140-171. Blackwell Publishing, Oxford.

Zagona, K. (2013). Tense, aspect and modality. M. Dikken (ed.), *Encyclopedia of Generative Syntax*, 746-792. Cambridge University Press., Cambridge, UK.

Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures

Bruno Gaume[1], Ludovic Tanguy[1], Cécile Fabre[1], Lydia-Mai Ho-Dac[1],
Bénédicte Pierrejean[1], Nabil Hathout[1], Jérôme Farinas[2], Julien Pinquier[2],
Lola Danet[3], Patrice Péran[4], Xavier de Boissezon[3], Mélanie Jucla[5]

[1] CLLE-ERSS: CNRS and University of Toulouse, Toulouse, France
{bruno.gaume, ludovic.tanguy, cecile.fabre, lydia-mai.ho-dac, benedicte.pierrejean, nabil.hathout}@univ-tlse2.fr

[2] IRIT: University of Toulouse and CNRS, Toulouse, France
{jerome.farinas, julien.pinquier}@irit.fr

[3] CHU de Toulouse & ToNIC: University of Toulouse, Inserm, UPS, Toulouse, France
lolla.danet@inserm.fr, deboissezon.xavier@chu-toulouse.fr

[4] ToNIC: University of Toulouse, Inserm, UPS, Toulouse, France
patrice.peran@inserm.fr

[5] URI Octogone-Lordat : Université de Toulouse, Toulouse, France
melanie.jucla@univ-tlse2.fr

Abstract. This paper is the fruit of a multidisciplinary project gathering researchers in Psycholinguistics, Neuropsychology, Computer Science, Natural Language Processing and Linguistics. It proposes a new data-based inductive method for automatically characterising the relation between pairs of words collected in psycholinguistics experiments on lexical access. This method takes advantage of four complementary computational measures of semantic similarity. We compare these techniques by assessing their correlation with a manual categorisation of 559 distinct word pairs, and with the distribution of data produced by 30 test subjects. We show that some measures are more correlated than others with the frequency of lexical associations, and that they also differ in the way they capture different semantic relations. This allows us to consider building a multidimensional lexical similarity to automate the classification of lexical associations.

1. Introduction

Assessing and characterising lexical access is one of the main interests of psycholinguists. To build their experimental material, psycholinguists frequently use measures obtained from the analysis of large corpora (see for instance lexical frequency measures; New et al., 2004). In order to specifically tackle lexical semantic relations, word association tasks are very useful tools. In such a task, a participant has to say (or write) a word in response to an auditory or written stimulus (e.g. “dog” as a response to “cat”). The variables typically analysed are latencies, error rate and the lexical frequency and length of the response. Of course, together with those quantitative data, a qualitative analysis of the grammatical and/or semantic relation between the stimulus and the answer helps addressing 1) how close two words can be in someone’s mental lexicon, 2) the nearest neighbours of a specific word 3) whether this network is affected by age (Burke and Peters, 1986), gender, sociodemographic status and language pathologies (Péran et al., 2004). However two main problems arise. First, we lack norms about the typical answers produced by a large sample of participants so that we cannot reliably know whether a stimulus/response pair is more or less plausible for a large number of words (see for French norms Alario & Ferrand, 1998 based on 300 words for young adults, de La Haye, 2003 based on 200 words for children and young adults and Tarrago et al., 2005 based on 150 words for elderly people). Second, a qualitative subject-by-subject and item-by-item analysis is time consuming and prone to interpretation. Such data can be obtained through the analysis of reference language data with Natural Language Processing (NLP) techniques and help psycholinguists to better and automatically analyse word association tasks.

In NLP, the use of data-based inductive methods for automatically measuring the similarity between words is one of the key task in computational semantics. If the first methods were based on the collocation frequency of words in large corpora (Church and Hanks 1990, Evert 2009), newer techniques rely on the principles of distributional semantics (Lenci 2008, Mikolov et al. 2013, inter alia). Nevertheless, even if the performance of these systems is sometimes impressive for some specific tasks (analogy resolution, lexical substitution, etc.), they usually fail to provide a fine grained characterisation of the relation between two words. Current distributional semantic models tend to aggregate all the classical lexical relations (e.g. synonymy, hypo/hypernymy, meronymy) and to confuse relations between similar words (e.g. *couch* - *sofa*) and relations between associated words (e.g. *couch* - *nap*). There is also need for evaluation data when comparing and assessing these techniques (Hill et al. 2005, Baroni and Lenci 2015).

This paper proposes a step toward the satisfaction of both needs. We use the data gathered in psycholinguistics experiments to compare different similarity measures and at the same time, investigate how using complementary computational semantic techniques can help characterising lexical relations between stimuli and responses provided by subjects in a word association task. The Evolex project (from which the data were collected) and the data collection process are both detailed in Section 2. Section 3 describes the manual annotation process and provides a linguistic analysis of the lexical relations in the dataset. We present the computational measures of semantic similarity in Section 4. Sections 5 contains the quantitative analyses and results while Section 6 presents a promising method able to characterise and cluster stimulus/response pairs.

2. Data collection: the Evolex protocol

2.1 Data collection

The Evolex Project is a multidisciplinary project gathering researchers in Psycholinguistics, Neuropsychology, Computer Science, NLP and Linguistics. Its main objective is threefold:

- to propose a new computerised tool to evaluate lexical access in population with or without language deficits;
- to complement and reinforce the neuropsychological characterisation of lexical access using a qualitative lexical analysis (and vice versa);
- to develop and train appropriated NLP tools to automatically measure and identify the lexical relations.

The Evolex protocol includes three different tasks to assess lexical access. The Verbal Fluency test is a common procedure that includes two semantic fluency tasks (Benton 1968) that consists in naming words belonging to the animal or fruit category and two phonemic fluency tasks (Newcombe 1969) that consist in naming words starting with the letters R or V. For the Picture Naming task: participants are shown a very explicit picture (e.g. *igloo*, *baby bottle*, *cat*) and have to vocalise the word depicted by the picture. The last task is the Word Association task. This paper focuses mainly on this task which consists in vocalising the first word coming to mind after listening to a simple word (e.g. *fruit*, *painting*, *igloo*).

The 60 items used as audio stimuli for the Word Association task were selected according to their grammatical category (nouns), number of syllables (the same amount of words of 1, 2 and 3 syllables) as well as their frequency in generic corpora (as given by the *Lexique* resource, New et al. 2004). A fixed order was defined i.e. the same sequence of items is given to all participants. We chose this parameter so that the inter subject discrepancy in the answers could not be attributed to a simple list order effect. To maximise the reproducibility of the experiment, the audio stimuli were produced by a speech synthesis tool[1]. The task aims at collecting data on natural lexical organisation. By asking the participants to respond as quickly as possible, the experimenter avoids their use of possible strategies. Response times were not used in the study presented here.

One of the key innovations of the Evolex protocol is to propose a computer-assisted method for collecting and processing the data. The software includes a system that automatically recognises and analyses the vocal response. An Automatic Speech Recognition (ASR) tool transcribed the response and recorded the reaction time (i.e. the time period between the beginning of the stimulus and the beginning of the subject answer).

[1] http://acapela-box.com/.

A web interface allows the user to correct the ASR transcription.

This paper exploits a first data set of pairs of words collected from a pilot study conducted with 30 participants with no language disorders, native French speaker, aged between 15 and 58 (mean age 31 ± 13.06), with variable levels of education (from 10 to 20 years of schooling, mean 15.4 ± 2.97). The following instructions were given to participants: "*You will hear French common nouns. You will have to pronounce the first word which comes to your mind related to the one you just heard as fast as possible. For instance, when you hear TABLE, you may answer CHAIR*".

2.2 Data preprocessing: cleaning up and normalisation

We collected the 1800 individual responses (30 subjects, 60 stimuli). We grouped and filtered them according to the following criteria: the response must be a monolexical noun or a proper name, in its non-inflected form. In addition, we rejected two stimuli (and their associated responses) because of a phonological confusion induced by the speech synthesis system. We used a combination of automated post-processing and final manual verification and ended up with a total of 1544 individual validated responses, corresponding to 559 distinct stimulus-response word pairs.

3. Qualitative analysis of data

The 559 distinct stimulus-response pairs have been annotated by two judges. The tagset is composed of 12 tags, as illustrated in Table 1, and covers four types of relations. The first type of relations are classical lexical relations, as found for example in the WordNet database (Fellbaum 1998). They include synonyms, antonyms, generic-specific relations (hypernyms, hyponyms, co-hyponyms and instance) and part-whole relations (meronyms and holonyms). A second type of relation called *associated* holds between words that are semantically related in a broader way: they tend to appear in the same contexts (both textual and referential) because they are connected within the same class of objects or events (Morris and Hirst 2004). Syntagmatic relations concern words that tend to combine to form larger syntactic constituents (expressions, compounds, etc.). In the example given in Table 1, *fleur* (flower) and *peau* (skin) are not semantically related, but they are associated in the expression à fleur de peau (hypersensitive, thin-skinned). The fourth relation (*phonology*) refers to a phonological proximity between words, with no semantic connection. For a small proportion of the pairs, no specific relation could be identified (*none found*) or the connection seemed too far-fetched (as in the example in the last line of Table 1).

Independent double annotation has been performed and followed by adjudication. After this first step, 69 out the 559 annotated pairs received more than one tag because they could be part of several relations. Another stage of collective adjudication has been carried out to retain only the relation that was considered most prominent, on the basis of priority rules. In the resulting dataset, classical lexical relations altogether represent almost half of the pairs (49.5%), among which co-hyponyms (sisters of the same superordinate) stand out, although all classical relations other than antonyms are also well represented. Associated pairs make up more than one third of the set (36.1%). Syntagmatic relations form 8.8% of the pairs. The amount of phonological links is almost negligible (0.9%).

Table 1. Breakdown of the semantic relations used to categorise the 559 distinct stimulus-response word pairs.

Relation	Example (stimulus / response)	# distinct pairs	%
antonym	*aube* (dawn) / *crépuscule* (dusk)	2	0.4%
associated	*balançoire* (swing) / *enfant* (child)	202	36.1%
co-hyponym	*balançoire* (swing) / *toboggan* (slide)	73	13.1%
holonym	*doigt* (finger) / *main* (hand)	29	5.2%
hypernym	*balançoire* (swing) / *jeu* (game)	52	9.3%
hyponym	*animal* (animal) / *chat* (cat)	45	8.1%
instance	*magicien* (wizard) / *Merlin* (Merlin)	6	1.1%
meronym	*balançoire* (swing) / *corde* (rope)	49	8.8%
phonology	*chapiteau* (circus tent) / *château* (castle)	5	0.9%
synonym	*canapé* (couch) / *sofa* (sofa)	21	3.8%

syntagmatic	*fleur* (flower) / *peau* (skin)	47	8.4%
none found	*perroquet* (parrot) / *placard* (closet)	28	5.0%

4. Computational measures of semantic similarity

In this section we describe the different techniques used in order to compute the similarity measures that we apply to the stimulus-response word pairs collected from the Word Association task. The four techniques we tested differ in two ways. First, different resources were used: the first two make use of a large corpus of French, FrWaC (Baroni et al. 2009) which is a collection of Web pages from the *.fr* domain and consists of 2 billion words. The latter techniques are based on the *TLF* (*Trésor de la Langue Française*, see Dendien and Pierrel, 2003) dictionary from which we extracted the full text of all the definitions. Both resources have been POS-tagged and lemmatised with the *Talismane* toolkit (Urieli 2013). The second dimension on which these techniques differ is the fact that we consider either the first order similarity (meaning that words collocated in a corpus are similar, and that dictionary headwords are similar to the words appearing in their definition) or second order similarity (also known as distributional similarity), considering that words sharing first-order similar words show a possibly different degree of similarity. Each of these techniques is described in the following subsections.

4.1 Corpus-based first-order similarity (collocates)

The simplest measure of similarity is to consider collocation, i.e. the fact that some words appear frequently and systematically together. This corpus-based measure has a large number of uses in NLP and corpus linguistics, and is known to capture a large variety of semantic relations (Evert 2009). It has also been shown to be correlated with words association data (Wettler et al. 2005).

We computed this similarity using Positive Pairwise Mutual Information, one of the most commonly used alternatives amongst collocation measures (Evert 2009). Each word was considered using its POS-tag and lemma, and its collocations were extracted in a symmetrical rectangular (unweighted) window of 3 words in both directions.

4.2 Corpus-based second-order similarity (word embeddings)

The second corpus-based measure relies on the principles of distributional semantics, which consider that words appearing in the same contexts have similar meanings. Second-order similarity can be computed in a number of ways (Baroni and Lenci 2010), but for a few years most of the work and research has focused on word embeddings. Word embeddings are a dense numeric representation of lexical items based on their distribution in a corpus. State-of-the-art methods for computing these embeddings rely on neural networks that are trained to predict words given context elements (or vice-versa), and are readily available. For this experiment, we used Word2vec (Mikolov et al. 2013), undoubtedly the most commonly used system and applied it to the same corpus used for first-order similarity i.e. FrWac. The following parameters were used: skipgram algorithm with negative sampling (rate 5), window size 5, 500 dimensions, subsampling rate 10^{-3}, 5 iterations, minimum frequency 100. As a result we obtained a dense matrix in which each word is represented by a numeric vector (of size 500). The cosine distance was then computed to measure the similarity between two words. Distributional semantics similarity measures are well known for capturing a wide spectrum of semantic relations (Baroni and Lenci 2011). This can be an issue for some tasks (Ferret, 2015) but was an asset in our case.

4.3 Dictionary-based first-order similarity (presence/absence in definitions)

The third technique, which uses a general-purpose dictionary for measuring first-order similarity, is based on a very simple principle: if a word appears in the definition of another word then the two words share a part of their semantic contents. We used this straightforward approach on the definitions extracted from the TLF dictionary without considering any explicit information that could be found in the dictionary such as cross-references. The only additional processing we applied was to symmetrise the relation. This similarity measure $s_{TLF}(x,y)$ $s_{TLF}(x,y)$ is therefore binary: the similarity between *x* and *y* is 1 if *x* appears in the TLF's definition of *y* or vice-versa (or both), and 0 otherwise.

4.4 Dictionary-based second-order similarity (random walk across definitions)

For second-order similarity we used a random walk approach (Bollobas, 2002). This graph traversal technique is used to define a broader, more robust[2] measure of similarity between the nodes of a graph. We applied this technique to the undirected, unweighted G_{TLF} G_{TLF} graph used in the first-order approach described in the previous subsection. Formally, this similarity measure is $P^t_{G_{TLF}}(x,y) \in [0,1]$ $P^t_{G_{TLF}}(x,y) \in [0,1]$, i.e. the probability of a walker crossing the links of G_{TLF} G_{TLF}, starting on vertex x x , to reach the vertex y y , after t t steps. This technique will therefore attribute a positive similarity score to two words whose definitions share words (the more words, and the more specific they are, the higher the value), or to two words appearing in the same definitions, and even to slightly more distant words in the original graphs. This method has proved to capture different kinds of semantic relation.

5. Quantitative analysis and results

As described in Sections 2 and 3, our dataset consists of 559 stimulus-response pairs of words, each with a hand-tagged semantic relation. In addition, we also know the response frequency, i.e. the number of subjects that gave the same response for a given stimulus as well as the four computed similarity values. We performed two kinds of analysis on this data.

First, we computed the correlation between the four similarity measures presented in Section 4 and the response frequency. We used the Spearman correlation coefficient over all pairs and obtained the scores presented in Table 2 below.

As can be seen, all correlation values are positive and statistically significant. The highest value is obtained for the dictionary-based second order similarity. For both resources, shifting from first to second order results in an increased correlation (up to 70% for Dictionary-based methods).

Table 2. Spearman correlation between similarity measures and the response frequency.

Similarity measure	Spearman's ρ	p-value
Corpus-based, 1st order	0.215	2.3e-07
Corpus-based, 2nd order	0.247	5.3e-09
Dictionary-based, 1st order	0.191	4.5e-06
Dictionary-based, 2nd order	0.325	1.7e-15

In order to get a more detailed view of the complementarity of these measures, and to examine the behaviour of these measures regarding the semantic relations between stimulus and response, we performed a multidimensional analysis. We ran a standard Principal Component Analysis on the matrix containing the similarity values and response frequency for each pair, and then projected the categories on the reduced vector space. The main factor map is presented in Figure 1 below, representing 66% of the global variance).

Several elements can be learned from this analysis. It clearly shows that the two resources (corpus and dictionary) provide different aspects of lexical similarity, and that the shifting from first to second order preserves these differences. When looking at the categorised semantic relations (cf. Section 3), several phenomena can be identified. First, it appears that all similarity measures are negatively correlated to non-classical relations. The none cases for which no semantic relation has been identified have low similarity values for all measures, and it is the same (to a lesser extent) for *phonology* and *instance* word pairs. Associated and syntagmatic relations appear in the centre of the factor map, indicating that no clear trend can be identified for these relations, although they are on the opposite side of the similarity vectors.

[2] We say "robust" in the sense that the similarities computed on graphs from different dictionaries are strongly related to one another, while this is not the case with first order methods (see Gaume et al., 2016).

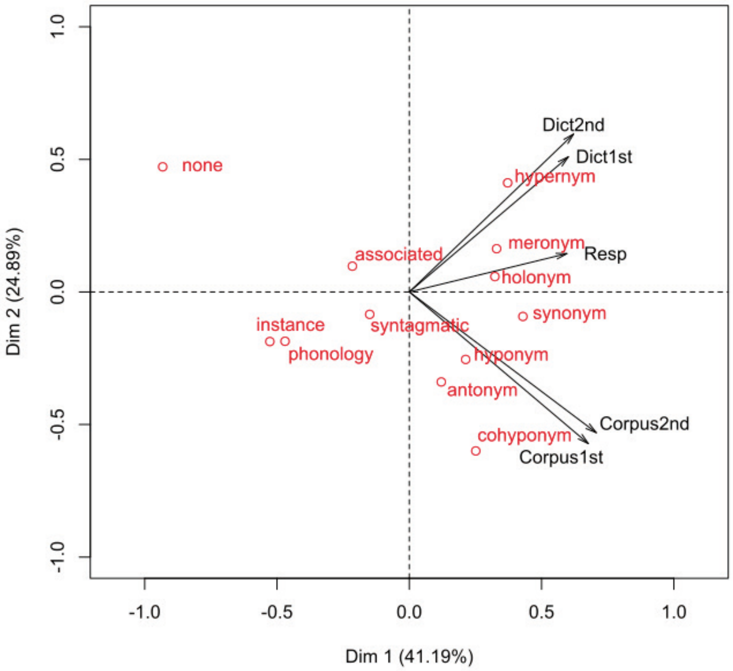


Figure 1. First factor map of PCA based on the 4 similarity measures and response frequency. Categorised relations (in red) are shown as additional variables.

This is somewhat surprising that even corpus-based first order similarity does not capture these cases. On the other (right) side of the map, we can find all classical semantic relations, although with varying correlations with the four similarity measures. It appears that dictionary-based methods capture the hypernymy relation more easily, while corpus-based methods favour co-hyponymy. Other relations are positively correlated with all measures, without a clear advantage for any of them. This indicates that automated measures can be useful for the detection of atypical responses. They will be tested against authenticated pathological responses in the near future.

In conclusion, we can see that the four tested methods manage to capture a significant part of the associations produced by subjects, with the more sophisticated (second order) measures showing a slightly higher correlation. The resources used for computing similarity have an influence on both the overall correlation, but more interestingly on the types of semantic relations between stimulus and response. However, none of these methods is particularly suited to identify non-paradigmatic associations.

6. Beyond semantic relations: clustering responses

Although the reliable identification of specific semantic relations between a stimulus and responses provided by the subjects is currently out of reach, some of the NLP techniques used to compute similarity can be used to provide a structure for the set of responses. This is especially the case for word embeddings, which are known to provide vector representation of words that are suitable for a number of semantic tasks.

For example, we can use these representations to identify clusters of responses based on their position in the vector space (vector space computed from the distribution of words in a corpus). We show here two examples of such analysis.

Focusing on the stimuli *igloo*” and “*cat*, we extracted for each one the word embeddings of all responses (and the stimulus) and represented them in a two-dimensional space by the means of a PCA on the initial 500-dimension vectors. The results can be seen in Figure 2 below. If the dimensions themselves cannot be interpreted, it appears that interesting clustering can be seen in the responses.

For *igloo*, we can see that all words related to the igloo's typical climate and environment are gathered close to the stimulus (*cold*, *ice*, *snow*), while the prototypical inhabitants (*Eskimo*, *Inuit*) and fauna (*penguin*, *walrus*) are farther on the left. The hypernym *house* is located in another area, this time closer to the top. Another interesting case in this example is the presence of *captain* in the responses: it refers to a fictional character named "*Captain Igloo*" who used to appear in TV commercials for frozen fish sticks. Its position in the figure is understandably the most extremely afar from the stimulus. It is important to note that the semantic relations of most of the responses with this stimulus fall under the "associated" category, with the exception of the meronym *ice*, the hypernym *house* and the syntagmatically-related *captain*. However, it appears that word embeddings are able to separate them efficiently in relevant subsets.

The results for *cat* are more self-explanatory, with the interesting case of *mouse* which is not considered as a close co-hyponym (as are *dog*, *rat* and *lion*) but more as an association because of the "cat and mouse" topoi.

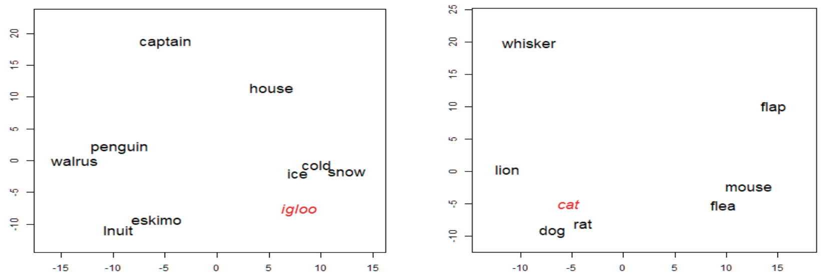


Figure 2. PCA maps of the responses to the stimuli (in red) "igloo" (left) and "cat" (right), based on word embeddings - manual translation to English.

7. Conclusion

In this paper, we described a series of experiments performed on the data collected from a word association task, in order to assess the possibility of using NLP techniques to automatically categorise the responses provided by non-pathological subjects. We manually tagged 559 different word pairs according to the lexical semantic relation between stimulus and response. We tested four different measures of similarity that differ on the resource used (a general corpus and a dictionary), and the nature of similarity (first and second-order). We showed that dictionary-based second-order similarity provides a measure that has the highest correlation with the data in terms of number of subjects who produced the responses. We also showed that if all of these different measures have very low scores for non-semantically related pairs, and favour some of the more classical relations (hypernymy, synonymy and co-hyponymy), they cannot be used without further development to identify the other relations, and especially the non-paradigmatic associations. However, we also showed that these techniques can prove surprisingly efficient for the clustering of responses according to stimulus-specific semantic relations.

There are other factors that need to be taken into account in the near future. The reaction time of each response is known to be a significant information for categorising subjects, as are of course the generic characteristics of the subjects (age, education level, etc.). But the most important perspective for the work and methods presented here is of course the analysis of data collected from pathological subjects. We are confident that the similarity measures will be able to identify non-typical responses, but we will need further analysis in order to associate the non-typicality with specific pathologies or disorders.

Acknowledgements

The authors wish to thank Britta Grusdt (formerly from the University of Osnabrück), Nicolas Viot, and all the students from the Toulouse school of speech therapy for their work in collecting and cleaning up the data.

References

Baroni, M., Bernardini, S., Ferraresi, A. & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, *43*(3), 209-226.

Baroni, M. & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673-721.

Benton, A.L. (1968). Differential behavioural effects in frontal lobe disease. *Neuropsychologia*, 6, 53-60.

Bollobas B. (2002). *Modern Graph Theory*, Springer-Verlag New York Inc.

Burke, D.M., & Peters, L. (1986). Word associations in old age: Evidence for consistency in semantic encoding during adulthood. *Psychology and Aging*, *1*(4), 283.

Church, K.W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

de La Haye, F. (2003). Normes d'associations verbales chez des enfants de 9, 10 et 11 ans et des adultes. *L`Année psychologique*, *103*(1), 109-130.

Dendien, J., & Pierrel, J.M. (2003). Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL*, *44*(2).

Evert, S. (2009). Corpora and collocations. A. Lüdeling, M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, Vol. 2. Berlin, New York, 1212-1248.

Ferrand, L., & Alario, F.X. (1998). Normes d'associations verbales pour 366 noms d'objets concrets. L'Année psychologique, *98*(4), 659-709.

Ferret O. (2015). Typing Relations in Distributional Thesauri. N. Gala, R. Rapp, G. Bel-Enguix (eds.). *Language Production, Cognition, and the Lexicon*. Springer.

Fellbaum, C. (ed., 1998). *WordNet: an Electronic Lexical Database*. MIT Press.

Gaume, B., Duvignau, K., Navarro, E., Desalle, Y., Cheung, H., Hsieh, S., Magistry, P., & Prévot, L. (2016). Skillex: a graph-based lexical score for measuring the semantic efficiency of used verbs by human subjects describing actions. *TAL*, 55.

Hill, F., Reichart, R. & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4).

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1-31.

Mikolov, T., Chen, K., Corrdao, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *ICLR Workshop*.

Morris, J., and Hirst, G. (2004). Non-classical lexical semantic relations. *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, 46-51.

New, B., Pallier, C., Brysbaert, M. & Ferrand L. (2004) Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments & Computers*, *36*(3).

Newcombe, F. (1969). *Missle Wounds of the Brain. A Study of Psychological Deficits*. London: Oxford University Press.

Péran, P., Démonet, J. F., Pernet, C., & Cardebat, D. (2004). Verb and noun generation tasks in Huntington's disease. *Movement Disorders*, *19*(5), 565-571.

Tarrago, R., Martin, S., De La Haye, F., & Brouillet, D. (2005). Normes d'associations verbales chez des sujets âgés. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *55*(4), 245-253.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis. University of Toulouse, France.

Wettler, M., Rapp, R. & Sedlmeier, P. (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, *12*(2-3).

A clustering approach for detecting defects in technical documents

Manel Mezghani[1], Juyeon Kang[1], Florence Sèdes[2]

[1] Prometil, 52 Rue Jacques Babinet, 31100 Toulouse, France
{m.mezghanni, j.kang}@prometil.com
[2] IRIT, University of Toulouse, CNRS, INPT, UPS, UT1, UT2J, France
florence.sedes@irit.fr

Abstract. Requirements are usually "hand-written" and suffers from several problems like redundancy and inconsistency. The problems of redundancy and inconsistency between requirements or sets of requirements impact negatively the success of final products. Manually processing these issues requires too much time and it is very costly. The main contribution of this paper is the use of k-means algorithm for a redundancy and inconsistency detection in a new context, which is Requirements Engineering context. Also, we introduce a pre-processing step based on the Natural Language Processing (NLP) techniques to see the impact of this latter to the k-means results. We use Part-Of-Speech (POS) tagging and noun chunking to detect technical business terms associated to the requirements documents that we analyze. We experiment this approach on real industrial datasets. The results show the efficiency of the k-means clustering algorithm especially with the pre-processing.

1. Introduction

For a system to become operational in real applications, several stages of conception, development, production, use, support and retirement must be followed (ISO/IEC TR 24748-1, 2010). During the conception stage, we identify and document the stakeholder's needs in the system requirements specification (Hull, 2011). Writing clearly all required elements without ambiguities (Berry, 2003) in the specifications is an essential task before passing to the development stage (Galin2003, Bourque2004). According to the 2015 Chaos report[1] by the Standish Group, only 29% of projects were successful[2], 50% of the challenged projects are related to the defects (Alshazly, 2014) from the Requirements Engineering (RE) and 70% of them come from the difficulties of understanding implicit requirements. All these defects do not lead to the failure but generate useless information. It is well known that the costs to fix defects increase much more after that the product is built than it would if the requirements defects were discovered during the requirements phase of a project (Glas, 2002; Stecklein, 2004).

When writing or revising a set of requirements, or any technical document, it is particularly challenging to make sure that texts are easily readable and are unambiguous for any domain actor. Experience shows that even with several levels of proofreading and validation, most texts still contain many language errors (lexical, grammatical, style), and a lack of overall concordance, or redundancy and inconsistency in the underlying meaning of requirements. Manually identifying redundant or inconsistent requirements is an obviously time-consuming and costly task. We focus in this paper on two critical issues in writing high quality requirements that can generate fatal defects in a product development stage: redundancy and inconsistency. We tackle these problems in terms of similarity between requirements since more than two similar requirements can be classified as redundant or inconsistent requirements.

[1] http://www.standishgroup.com.

[2] They studied 50,000 projects around the world, ranging from tiny enhancements to massive systems re-engineering implementations.

The problems of redundancy and inconsistency can be handled according to different technologies. We focus on artificial intelligence approaches and more precisely classification approaches. Automatic classification of requirements is widely used in the literature using Convolutional Neural Networks (Winkler, 2016), Naives Bayes classifier (Knauss, 2012), text classification algorithms (Ott, 2013). Data classification approaches could be data clustering through algorithm such as K-means. This latter is studied in different contexts due to its efficiency (Jain, 2010). However, in requirements engineering context, we could not find advanced works on the redundancy and inconsistency issues using k-means algorithm.

The main contribution of this paper is the use of k-means algorithm for a redundancy and inconsistency detection in a new context, which is requirements engineering context. Also, we introduce a pre-processing step based on the Natural Language Processing (NLP) techniques to assess the impact of this latter to the k-means results. We use Part-Of-Speech (POS) tagging and noun chunking to detect technical business terms associated to the requirements documents that we analyze.

This paper is structured as follows: In section 2, we present related works on the redundancy and inconsistency detection through artificial intelligence approach and especially k-means technique. In section 3, we present the datasets used for the experimental part. In section 4, we present our clustering approach. In section 5, we discuss the associated results. In section 6, we conclude and give some future research directions.

2. Related works

In this section, we first present related works associated to redundancy and inconsistency detection in specifications documents or technical documents. Second, we give some researches focusing on text pre-processing in requirements engineering context. Finally, we focus on approaches using k-means clustering in the latter context.

2.1 Redundancy and inconsistency detection

Researches on redundancy detection began by traditional bag-of-words (BOW), TF-IDF frequency matrix, and n-gram language modelling (Allan, 2000) (Brown, 1992). Then, researchers like Juergens et al. (Juergens, 2010) use ConQAT to identity copy-and-paste reuses in requirements specifications. Falessi et al. (Falessi, 2013) detect similar content using information retrieval methods such as Latent Semantic Analysis. They compare NLP techniques on a given dataset to correctly identify equivalent requirements. Rago et al. (Rago, 2016) extend the work presented in (Falessi, 2013) specifically for use cases. Their tool, ReqAlign, combines several text processing techniques such as a use case-aware classifier and a customized algorithm for sequence alignment.

Inconsistency is analyzed in (Belsis, 2014) by proposing the framework of a patterns-based unsupervised requirements clustering (based on k-means algorithm), called PBURC, which makes use of machine-learning methods for requirements validation. This approach aims to overcome data inconsistencies and effectively determine appropriate requirements clusters for optimal definition of software development sprints. Dermeval et al., (Dermeval, 2016) present a survey about how using ontologies in RE activities both in industry and academy, is beneficial, especially for reducing ambiguity, inconsistency and incompleteness of requirements.

2.2 Pre-processing

Some researches introduce pre-processing steps in requirements analysis context. According to (Abad, 2017), the pre-processing helps reducing the inconsistency of requirements specifications by leveraging rich sentence features and latent co-occurrence relations. It is applied through: i) a Part-Of-Speech tagger (Klein, 2003), ii) an entity tagging through a supervised training data, iii) a temporal tagging through a rule-based temporal tagger and iv) co-occurrence counts and regular expressions. This pre-processing approach improved the performance of an existing classification method.

Pre-processing data for redundancy detection is used in (Fu, 2017) by performing standard NLP techniques such as removing English stop words and striping off the newsgroup related meta-data (including noisy headers, footers and quotes). The Joint Neural Network for redundancy detection approach in (Fu, 2017) also uses normalized bag-of-words (BOW) as a pre-processing approach. The normalized BOW generates a global unigram based dictionary mapping. With the presence of the uni-gram indexer, the authors could readily remove low frequency terms and lengthy snippets.

2.3 K-means

K-means clustering is a type of unsupervised learning approach, which is used on unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to cluster the data into k groups (k number of groups).

Classifying requirements is an important task in requirements engineering. Recently, some studies introduce k-means in requirements classification tasks. Notably, (Abad, 2017) applies different approaches such as i) topic modelling using Latent Dirichlet Allocation (LDA) and Biterm Topic Model (BTM) and ii) clustering using K-means, Hierarchical approach and Hybrid (k-means and hierarchical) to classify requirements into functional (FR) and non-functional requirements (NFR).

3. Datasets

To test our approach, we extracted requirements from 22 industrial specifications (~2000 pages). From this, we constructed three different datasets (corpus1, corpus2 and corpus3) explained below. For confidentiality issues, we are not allowed to reveal the identity of the companies. The main features considered to validate our datasets are: 1) texts following various kinds of business style and format guidelines imposed by companies, 2) texts coming from various industrial areas: aeronautic, automobile, spatial, telecommunication, finance, energy. These datasets enable us to analyze different types of redundancy and inconsistency in terms of frequency and context. We present characteristics of these datasets (written in English) as follows:

- Corpus1: dataset that contains 38 requirements fully redundant according to our expert,
- Corpus2: dataset that contains 42 requirements fully inconsistent according to our expert,
- Corpus3: dataset that contains 337 requirements randomly chosen with no a priori information of redundancy and inconsistency,

The expert in this work means requirements engineer with more than 15 years of experiences (industrial and academic).

4. Clustering Approach

In this section, we present the basics of k-means clustering algorithm and the results of our approach. We also analyze the impact of the pre-processing step on the results.

4.1 K-means algorithm

The k-means algorithm is used to partition a given set of observations into a predefined amount of *k* clusters. K-means algorithm is a very popular approach due to its efficiency. However, it needs a predefined value of K as an input, which is the main issue about using this algorithm.

Some researchers focus on this issue and present solutions based on the graphical (e.g. elbow approach, silhouette and Inertia or numerical value (e.g. statistic gap (Mohajer, 2010)). We use in this paper the following solutions to calculate the value of K:

- Inertia: calculated as the sum of squared distance for each point to its closest centroid, i.e., its assigned cluster. It can be recognized as a measure of how internally coherent clusters are.
- Statistic gap: calculates a goodness of clustering measure. The statistic gap standardizes the graph of $log(W_k)$, where W_k is the within-cluster dispersion, by comparing it to its expectation under an appropriate null reference distribution of the data (Mohajer, 2010).

4.2 Determining the best number of K

We apply the k-means algorithm on the datasets already detailed in section 3. To choose the best similarity metric, we tested different similarity metrics such as Euclidean distance, TF-IDF, JACCARD, Correlation and Dice. K-means is applied using the Euclidean distance as similarity metric since we had best results comparing to other similarity metrics according to our expert.

We first begin by determining the best number of K by calculating the inertia in figure 2.

According to figure 2, determining visually the number of k cannot always be unambiguously identified. We can estimate the number of K between 25 and 30. To leverage this ambiguity, we choose to apply the statistic gap approach which allows to obtain a numerical value reflecting the coherence of the clusters.

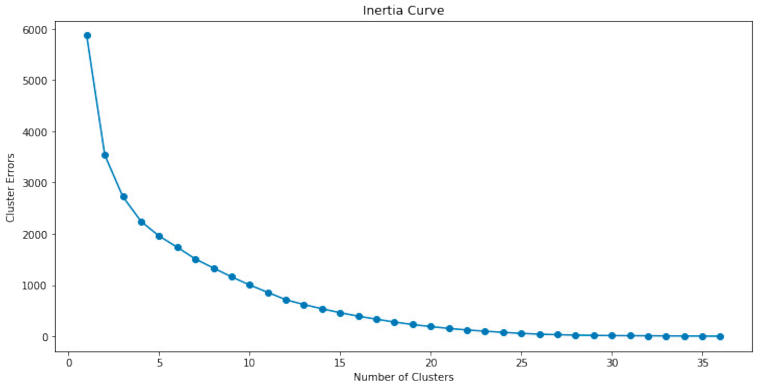


Fig. 2. Inertia curve for Corpus1 dataset

We apply the statistic gap to our datasets and the best number of k is presented in Table 1.

4.3 Validation approach

Since we use an unsupervised clustering approach, we do not have any ground truth about the redundancy and/or the inconsistency of the requirements. So, we give the results related to the best value of k to our domain expert in order that the expert evaluates the relevance of the generated clusters. A cluster may contain one or more requirement(s).

For a given k value, the validation is done according to two methods:

- "Strict" validation (SV): we assume that a relevant cluster contains 100\% correct requirements (fully redundant or incoherent requirements), which means that we discard clusters with partially relevant requirements. Also, we consider only clusters with more than one requirement.
- "Average" validation (AV): we calculate the average of relevant requirements per cluster.

$$\boxed{AV_k = (\Sigma i=1, k\ \mathit{precision}(C_i))/\ k' \ (1)}$$

where AV_k is the average validation for a given value of k. k is the number of clusters. k' is the number of clusters which their number of requirements is >1. i ∈{1, k} is the value of k and $\mathit{precision}(c_i)$ is defined as:

$$\mathit{precision}(C_i)= NumberofRelevantRequirements/TotalNumberofRequirements \qquad (2)$$

4.5 Classification results with the pre-processing step

For the pre-processing step, we use the Part-Of-Speech (POS) tagging and Noun chunking from SpaCy[3] as a popular tool in natural language processing field. SpaCy is a free open-source library featuring state-of-the-art speed and accuracy and a powerful Python API.

After applying this tagging approach, we proceed to detect technical terms according to some combination of tags.

According to our RE expert, technical business terms are often expressed in open or hyphenated compound words (e.g. *high speed*, *safety-critical*) and we observe that they are always parts of a noun chunk[4]. For this paper, we first extracted all noun chunks from our Corpus1, then observed the syntactic patterns inside noun chunks referring to POStags, obtained by SpaCy. The most used 13 combination patterns in business terms are selected and validated in collaboration with our RE expert: for example, noun-noun (e.g. *runway overrun*), adjective-noun (e.g. *normal mode*), proer_noun-noun (e.g. *BSP data*), adjective-adjective-noun (e.g. *amber visual indication*), noun-noun-noun (e.g. *output voltage value*).

So, we apply the k-means algorithm on dataset containing technical terms to see the impact of this pre-processing on the results. Table 2 summarizes the different results obtained from the same experiments presented in 4.3.

[3] https://spacy.io/.

[4] A noun chunk is a noun plus the words describing the noun.

Table 1. Results: Best value of K, validation results and the associated number of relevant clusters for each dataset.

Dataset	Best value of K	SV (Nb. of relevant clusters)	AV (Nb. of relevant clusters)
Corpus1	30	100% (8)	100% (8)
Corpus2	17	100% (15)	100% (15)
Corpus3	26	22% (4)	30.96% (18)

Table 2. Results with pre-processing: Best value of K, validation results and the associated number of relevant clusters for each dataset.

Dataset	Best value of K	SV (Nb. of relevant clusters)	AV (Nb. of relevant clusters)
Corpus1	28	100% (10)	100% (10)
Corpus2	24	92.85% (13)	92.85% (13)
Corpus3	36	22.22% (6)	39.20% (27)

In Corpus1 dataset, we have 100% of relevant clusters and 10 relevant clusters for which the number of requirements is greater than 1 in both validations. The clustering has detected clusters with the right redundancy information but two more relevant clusters than the clustering without preprocessing. In this case, the tagging has shown its efficiency to improve redundancy detection results.

In Corpus2 dataset, we have 92.85% of relevant clusters and 13 relevant clusters which their number of requirements is greater than 1 in both validations. The clustering has detected clusters with the right inconsistency information but two less relevant clusters than the clustering without preprocessing. In this case, the preprocessing has shown its inefficiency to improve inconsistency detection results.

In Corpus3 dataset, we have the same relevant value of the strict validation comparing to the Table 1. However, the number of relevant clusters is higher. For the average validation, we clearly see an improvement of the percentage of relevant clusters and the total number of relevant clusters. The preprocessing has improved the rate of the redundancy/inconsistency detection.

5. Discussion

The k-means results are given to our domain expert to judge the best value of k from his/her own domain-based expertise. We found a difference between the generated k value (according to the statistic gap) and the best value according to our expert.

For the results without pre-processing, the results are as follows: for to Corpus1, our expert assume that 23 (instead of 30) is the best value of k with 100% of relevance (for SV) and with 13 relevant clusters (instead of 8). For Corpus2, our expert assumes that 18 (instead of 17) is the best value of k with 100% of relevance (for SV) and with 16 relevant clusters (instead of 15).

For the results with pre-processing, the results are as follows: for to Corpus1, our expert assume that 23 (instead of 28) is the best value of k with 100% of relevance (for SV) and with 14 relevant clusters (instead of 13). For Corpus2, our expert assumes that 25 (instead of 24) is the best value of k with 100% of relevance (for SV) and with 15 relevant clusters (instead of 13).

The nature of the dataset is very important to define the best value of k. In fact, if the dataset contains too many identical requirements, the k-means algorithm tends to cluster these requirements together and then very similar requirements will be discarded from these clusters and putted in other clusters. This is the case for Corpus3 where we found repetitions of many requirements. We should then take into consideration this characteristic to analyze more efficiently specifications.

Also, we experiment this clustering approach on large dataset (not mentioned above) with ~900 requirements. This type of dataset is very noisy with many identical requirements belonging to different chapters or sections. The results on this dataset are not satisfying. Taking into consideration the information about document hierarchy could help us to analyze different sub-documents and then shorter datasets to find the best clusters.

6. Conclusion

In this paper, we proposed an automatic approach for redundancy and inconsistency detection in requirements engineering context. This approach is based on an artificial intelligence technique and more precisely unsupervised machine learning algorithm, k-means. This approach is tested on real industrial datasets with different characteristics of redundancy and/or inconsistency. Also, we introduced the pre-processing step based on the NLP pre-processing to see the impact of this latter to the k-means results. We used Part-Of-Speech (POS) tagging and noun chunking to detect technical business terms associated to the requirements documents that we analyze.

K-means algorithm is tested according to the best k value generated by the statistic gap method. According to Corpus1 (redundant) and Corpus2 (inconsistent), k-means provides very relevant results by providing only clusters (with more than one requirement) with relevant information. Pre-processing has improved the rate of redundancy detection but not the rate of the inconsistency detection. According to Corpus3, the results show the importance of the pre-processing step to improve the clustering results in terms of precision and the number of detected clusters.

Even with high quality results for Corpus1 and Corpus2, we are not able yet to differentiate redundancy or inconsistency in very similar clusters in Corpus3. To overcome this shortcoming, we plan to apply another clustering approach on similar clusters. This new clustering will be based on semantic features. Also, we plan to eliminate identical requirements belonging to the same chapter before applying clustering to improve clustering. After improvements, this work will be integrated in the industrial tool: *Semios for requirements*[5].

Acknowledgements

This work is financially supported by the Occitanie region of France in the framework of CLE (Contrat de recherche Laboratoires-Entreprises)-ELENAA (des Exigences en LanguEs Naturelles à leurs Analyses Automatiques) project.

References

Dermeval, D. (2016). Applications of ontologies in requirements engineering: a systematic review of the literature. *Requirements Engineering*, *21*(4), 405-437.

Frenay, B. & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. IEEE *Transactions on Neural Networks and Learning Systems*, *25*(5), 845-869.

Belsis, P., Koutoumanos, A. & Sgouropoulou, C. (2014). PBURC: a patterns-based, unsupervised requirements clustering framework for distributed agile software development. *Requirements Engineering*, *19*(2), 213-225.

Anil, K. Jain (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666.

Rago, A., Marcos, C. & Diaz-Pace, J.A. (2016). Identifying duplicate functionality in textual use cases by aligning semantic actions. *Software & Systems Modeling*, *15*(2), 579-603.

Falessi, D., Cantone, G. & Canfora, G. (2013). Empirical Principles and an Industrial Case Study in Retrieving Equivalent Requirements via Natural Language Processing Techniques. *IEEE Trans. Softw. Eng.*, *39*(1), 18-44.

Juergens, E., Deissenboeck, F., Feilkas, M., Hummel, B., Schaetz, B., Wagner, S., Domann, C. & Streit, J. (2010). Can Clone Detection Support Quality Assessments of Requirements Specifications?. *Proceedings of the 32Nd ACM/IEEE International Conference on Software Engineering*,Volume 2, 79-88.

Fu, X., Ch'ng, E., Aickelin, U. & See, S. (2017). 2017 CRNN: A Joint Neural Network for Redundancy Detection. *IEEE International Conference on Smart Computing (SMARTCOMP)*, 1-8.

Abad, Z.S.H., Karras, O., Ghazi, P., Glinz, M., Ruhe, G. & Schneider, K. (2017). What Works Better? A Study of Classifying Requirements. *IEEE 25th International Requirements Engineering Conference (RE)*, 496-501.

[5] http://www.semiosapp.com/index.php?lang=en.

Klein, D. & Manning, C.D. (2003). Accurate Unlexicalized Parsing. In: *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Volume 1, 423-430.

MacQueen, J. (1967). Some method for classification and analysis fo multivariante observations. L.M. Le Cam, J. Neyman, J. (eds.), *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, University of California Press, 281-297.

Winkler, J. & Vogelsang, A. (2016). Automatic Classification of Requirements Based on Convolutional Neural Networks. *IEEE 24th International Requirements Engineering Conference Workshops (REW)*, 39-45.

Ott, D. (2013). Automatic Requirement Categorization of Large Natural Language Specifications at Mercedes-Benz for Review Improvements. *Requirements Engineering: Foundation for Software Quality: 19th International Working Conference, REFSQ 2013*, Essen, 50-64.

Knauss, E., Damian, D., Poo-Caamaño, G. & Cleland-Huang, J. (2012). Detecting and classifying patterns of requirements clarifications. In: *20th IEEE International Requirements Engineering Conference (RE)*, 251-260.

Allan, J., Lavrenko, V., Malin, D. & Swan, R. (2000). Detections, bounds and timelines: Umass and tdt-3. *Proceedings of Topic Detection and Tracking Workshop (TDT-3)*, Vienna, VA, 167-174.

Brown, P.F., deSouza, P.V., Mercer, R.L., Watson, T.J., Pietra, V.J.D. and Lai, J.C. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, *18*(4), 467-480.

Mohajer, M., Englmeier, K.-H. & Schmid, V.J. (2010). A comparison of Gap statistic definitions with and without logarithm function, Vol 96, http://nbn-resolving.de/urn/resolver.pl?urn=nbn:de:bvb:19-epub-11920-3.

Hull, E., Jackson, K. & Dick, J. (2011). *Requirements Engineering*. Springer-Verlag London.

Glas, R.L. (2002). *Facts and Fallacies of Software Engineering*. Addison-Wesley Professional.

Stecklein, J.M., Dabney, J., Dick, B., Haskins, B., Lovell, R. & Moroney, G. (2004). Error Cost Escalation Through the Project Life Cycle. *Proceedings of the 14th Annual International Symposium*, Toulouse, France.

Galin, D. (2003). *Software Quality Assurance: From Theory to Implementation*. Pearson.

Bourque, P. (2004). *Guide to the Software Engineering Body of Knowledge (SWEBOK) Guide*. IEEE Computer Society.

Berry, D.M., Kamsties, E. & Krieger, M.M. (2003). *From Contract Drafting to Software Specification: Linguistic Sources of Ambiguity*. https://fr.scribd.com/document/76672102/Ambiguity-Handbook.

Alshazly, A.A., Elfatatry, A.M., Abougabal, M.S., (2014). *Detecting defects in software requirements specification*, Alexandria Engineering Journal, *53*(3), 513-527.

Using smog-related data of Chinese Sina Weibo to explore correlation between health issues and relevant regions

Qinran Dang[1], Nicolas Turenne[2], Mathieu Valette[3]

[1] ERTIM, Institut National des Langues et Civilisations orientales, 2 rue de Lille, Paris
{qinran.dang}@inalco.fr
[2] LISIS, Université Paris-Est Marne-la-Vallée, 5 Boulevard Descartes, 77420 Champs-sur-Marne
{nturenne}@u-pem.fr
[3] ERTIM, Institut National des Langues et Civilisations orientales, 2 rue de Lille, Paris
{mvalette}@inalco.fr

Abstract. Sina Weibo is an important platform of information spreading and discussion with more than 1,17 hundred million productions of messages per day. Since 2008, air pollution has become a serious concern in China and well discussed on Weibo. In this article we will propose a multidisciplinary methodological framework to investigate 1) which public health issues are discussed in the smog-related weibo; 2) how these health issues are distributed geographically; 3) are there some interactive correlations between a relevant region and a certain health issues? Our methodology based on three complementary statistic models contributes to proceed a series of study: lexical, discourse and geographic analysis. Finally, we had obtained 7 distinct types of health issues intensively located in Beijing, Shanghai, Hebei, Tianjin. This study proves that there exists an interactive correlation between a certain type of health issues and its geographic location in regard to the smog problems in China.

1. Introduction

With the continuous development of computer technology, Sina Weibo has influenced the life of Chinese people in all aspects and has become an important platform on which to share information. As a free social network, Sina Weibo produces more than 117 millions of messages[1] per day coming from 7.5 million users[2]. These messages, limited to 140-characters, are called Weibo ("microblog"). Generally, Weibo presents either current/hot affairs, comments on current events, or discussion of hot topics. Compared to traditional media, furthermore, the accessible convenience and real-time features of Weibo make it possible to rapidly publish and update daily information. Therefore, Sina Weibo has gradually become an important source of real-time information and a vehicle of communication for the Chinese public.

The mining and analysis of the related-top-topics on Weibo would help us to explore and grasp the evolution of audiences' feelings, concerns and opinions about current events. As we know, since 2008, air pollution has become a serious concern for Chinese people. Three years later, this problem took on an unprecedented scale, and a dedicated name – 雾霾 (wù maí), i.e. 'smog' – appeared in the Chinese media and was circulated on social networks. Thus, in this article our aim is to investigate 1) which public health issues are discussed in the smog-related Weibo corpus; 2) how these health issues are distributed geographically by province; and 3) are there some interactive correlations between a specific region and certain characteristic health issues?

[1] Reference 'Q2 statistics and information graph of Weibo 2012': https://www.zhihu.com/question/20267900.

[2] Reference 'Report of development of users in Weibo 2016': http://www.wesdom.me/Weibo-pdf/.

2. Related works

The research field in Chinese social media has focused on hot events' detection (Yang 2013), sentiment analysis studies (Cui 2013, Shi 2012), or analysis of opinion (Wu 2014), or with other massive researches on health issues with Twitter data – such as the use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic (Signorini 2011), the use of Hangeul twitter to track and predict human influenza infection (Kim 2011) and the analysis of microblog discourse on HIV prevention in Vietnam (Ho-Dinh 2014). By comparison, there have been few studies exploring health issues (Wang 2014) in terms of/regarding microblog in China. Thus, our article is the first multidisciplinary exploration of health issues exposed by the smog-related corpus of Chinese social media – WEIBO.

3. Corpus and Data Exploration

3.1 Corpus construction and cleaning

In order to build our corpus, we collected and stored a large sample of public Weibo from the Beijing Language and Culture University Corpus Centre (BCC[3]) – an online data system with a size of about ten billion, who created a large full-text retrieval corpus with multiple languages, including Chinese (Xun 2016). Thanks to the multiple supports of search queries (character-based, word-based, wildcards, splittable words, and wildcards+any of those three expression types), we have built our own list of search queries, which matches a set of smog-related terms, such as 雾霾 (smog), 大雾 (fog), 霾 (smog), 空气污染 (air pollution), 大气污染 (atmospheric pollution), PM2.5 (*Particular Matter Ø* 2.5 µm), PM10(*Particular Matter Ø 10* µm). Finally, we have collected a total of 18,898 Weibo messages with 64,098 words after filtration and removal of duplicates.

It should be noted that with the original corpus of BCC we obtained 85% proper Weibo, the remaining 15% being 'dirty' messages which contain mainly two types of duplicates to be filtered and three types of noises to clean.

Here are two examples of duplicate Weibo:

- Two messages share not only exactly the same words, but also the same number of words, the only difference is hidden in invisible tokens, such as blank space or tabulation.

e.g. Begin to like windy days since there are smog days.

○ 自打有了雾霾天，就爱上了大风天

○ 自打有了雾霾天，就爱上了大风天

- Two messages have about 5% divergences in term of different words.

e.g. Hazardous substances of smog concerns mainly aerosol particles, which can cause rhinitis, bronchitis and other diseases, serious smog will even cause cardiopathy, particulate pollutants may incur myocardial infarction.

○ 雾霾中对健康有害主要是气溶胶粒子, 易引起鼻炎, 支气管炎等病症, 极重度雾霾还是心脏杀手, 颗粒污染物可能引发心肌梗死。雾霾

○ 雾霾中对健康有害主要是气溶胶粒子, 易引起鼻炎, 支气管炎等病症, 极重度雾霾还是心脏杀手, 颗粒污染物可能引发心肌梗死。雾霾天

The filtering of Weibo consists mainly of 'noise reduction' referring to:

- removing non-ASCII Weibo or unconvertible information in Weibo: punctuation, numbers, characters, emoticons, etc.;
- delete informal messaging conventions, e.g. email with @, links with www or http//, etc.
- remove manually the ambiguous Weibo messages, especially the messages with keyword 大雾 (fog), because this word is used to create metaphors:

e.g. The eyes are no longer the window of the soul because of the fog. Life also has many fog moments!

○ 大雾弥漫, 眼睛便不再是心灵之窗。人生同样有很多次的大雾 ！

[3] Available on http://bcc.blcu.edu.cn/.

3.2 Corpus words segmentation and part-of-speech(POS) tagging

After all these corpus cleaning manipulations, our next step aims at word segmentation and word POS tagging. As we all know, Chinese is a continuous language, which requires a pre-treatment of semantic segmentation to identify each word. Therefore, we employed a Python Chinese word segmentation module named 'Jieba' [4](Chinese for 'to stutter'). This can cut a sentence into separate semantic words in order to meet our text analysis requirements. The algorithm implemented into Jieba is based on a prefix dictionary structure to achieve efficient word graph scanning. Jieba had built a directed acyclic graph (DAG) for all possible word combinations, and used dynamic programming to find the most probable combination based on the word frequency. For unknown words, an HMM-based model is used with the Viterbi algorithm. Since one of our objectives was to study whether there is a correlation between health issues discussed in the smog-related messages in Weibo and their adjoining geological distribution, we needed to find all the three categories' keywords that appeared in the Weibo messages and to mark them by a POS tagging. Thanks to the import-dictionary system of Jieba, we have added respectively three dictionaries[5] to tag the three categories of keywords:

Table 1. Three categories of POS-tagging.

Keywords category	POS tagging composition	Example(keywords)
health-relate/ medicine-related	disease: word+disease medical terminology: word+Dns[1]/Dv	肺炎 disease (pneumonia) 呼吸道 Dns (respiratory tract) 看病 Dv (see a doctor)
city/province	city: word+city province: word+province	南京 city (nanjing) 河北 province (hebei)
denominations of smog	smog: word+denowumai fog: word+denowu haze: word+denomai air pollution: word+denopollu Particular matter : word+denopm	雾霾denowumai (smog) 大雾 denowu (fog) 霾 denomai (smog) 空气污染 *kongqiwuran* (air pollution) PM2.5denopm (*Particular matter Ø 2.5* µm)

3.3 Corpus organization by regions

In an attempt to explore the correlations between health issues and relevant regions in which a certain disease or symptoms are specific, besides words POS tagging work, we had to reorganize our corpus by region.[6] To do this, we adopted a simple method which located each Weibo in its geographic position (city/province of belonging): first of all, we listed our cities and provinces; then, we picked up one target city/province as a search term and extracted all the Weibo that contained the keyword; after that, we classified these selected Weibo into their city/province family and tagged them by a determinate metadata[7] "<city=xxx>" or "<province=xxx>". This procedure continues until all the corpus has been browsed. In addition, we structured our Weibo corpus with three other different types of metadata (Table 2): type of discourse <nature = message>; site <site = weibo>, and id<id = 1>. This metadata information helped us to multiply the exploration of our corpora.

[4] 'Jieba' (Chinese for 'to stutter') Chinese text segmentation: built to be the best Python Chinese word segmentation module. Available at https://github.com/fxsjy/jieba.

[5] We have three categories of dictionaries: a dictionary of the division of administrative regions of China (116,402 entries); the dictionary of health issues (31,295 entries); the dictionary of smog-related terms (60 entries).

[6] In order to simplify the denomination of each region, in the definition of 'region' in our corpus, we mix 23 provinces called 'province', 4 municipalities and 5 autonomous regions called 'city'.

[7] For example: if in a Weibo message we meet a term like 北京 (Beijing), we will categorize this Weibo into 'group city Beijing', which is given a determinate metadata tagged by "< >", in this case, '<city=Beijing>'. This action will be repeated throughout the corpus until it finds all the determinate city-terms or province-terms.

Table 2: Corpus metadata information

Corpus	Nb of Weibo	Nb of city/ region	Nb of forms	Nb of words	Nb of hapax	Metadata type
Weibo	18,898		798,162	64,098	32,324	3
Weibo- city	5631	24	219,328	21,346	9,879	4
Weibo- region	3849	32	151,696	18,370	9394	4

3.4 Selection of investigative regions

Thanks to a graphical mapping tool DITUHUI[8] developed by Supermap – provider of GIS (Geographic Information System) software products and services in Asia –we have obtained this cartography showing the frequency of region occurrences in our smog-related corpus.

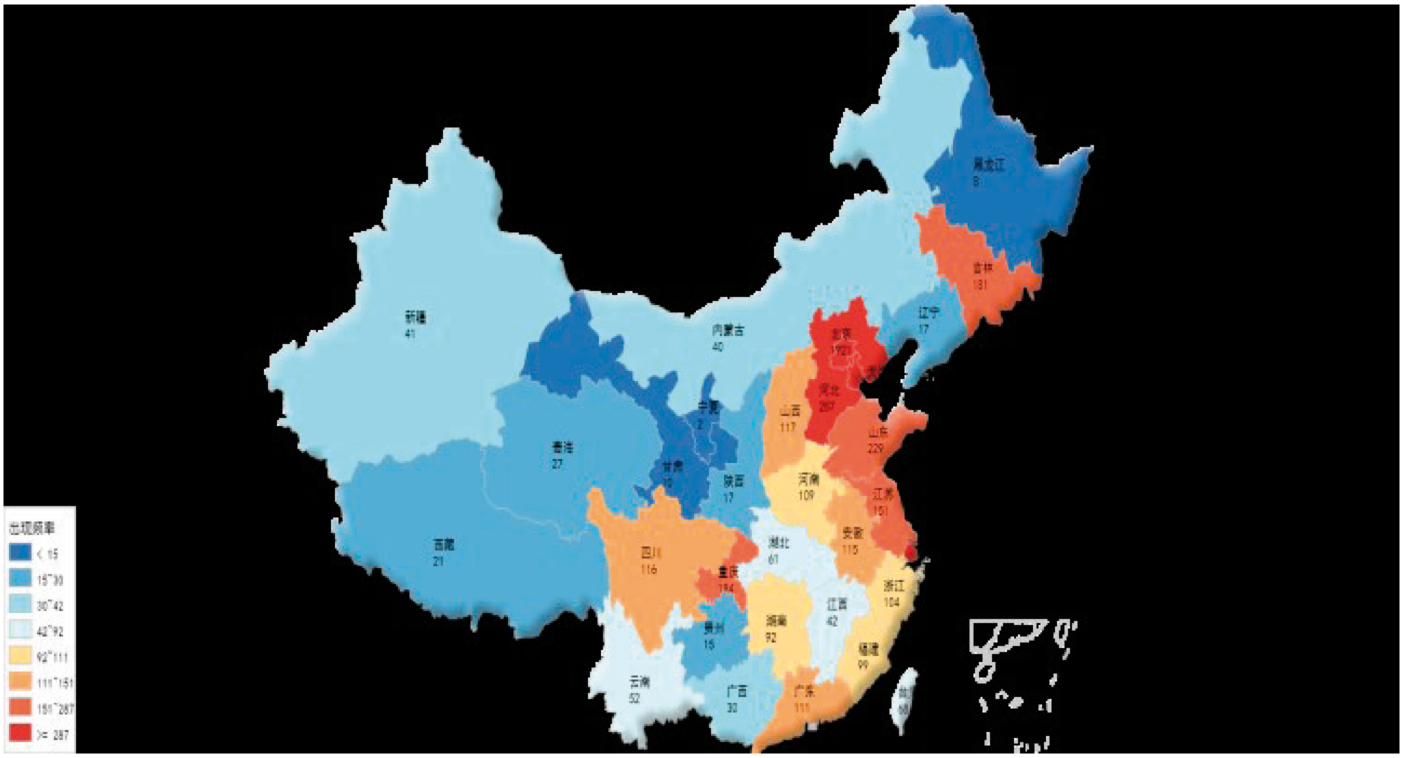


Fig. 1. Frequency of region occurrences in smog-related Weibo in China.

As we can observe through this graph (Fig.1), smog-related Weibo spread all over China and disperse in every region. The colour of an area reflects how many times its name has been mentioned in our corpus. The closer the colour is to dark blue (i.e. the lowest degree on the frequency scale), the lower the frequency of this province's occurrence; in other words, those regions coloured in dark blue are discussed very little when we talk about air pollution in China; however, if the colour is closer to dark red – the other end of the scale – these regions are much more recurrent in our smog-related discussion messages.

Figure 1 shows us a global visualization of the geographic distribution of smog-related Weibo across China. Obviously, 北京 beijing, 上海 shanghai, 天津 tianjin and 河北 hebei rank among the most discussed regions in our corpus. As the political and cultural centre, 北京 beijing has been mentioned 1921 times and 上海 shanghai 699 times. The frequency of occurrence of 天津 tianjin and 河北 hebei is respectively 390 times and 297 times. Indeed, in consideration of the actual conditions, the air is seriously polluted in these regions, especially in Beijing, Hebei and Tianjin in the north, but as one of the most developed southern cities, Shanghai is also inevitably affected by air pollution. Thus, these four regions are considered as our subjects of investigation in terms of geographic location to explore relevant health issues.

[8] https://www.supermap.com/en/html/SuperMap_GIS_Company.html.

3.5 Health issues exposed in the corpus

With health-related/medicine-related keywords POS tagging, we have collected a list of relevant health issues presented in Table 3:

Table 3: Health issues exposed in weibo corpus.

Disease	过敏 allergy, 癌症 cancer, 感冒 influenza, 肺癌 lung cancer, 支气管炎 bronchitis, 肺炎 pneumonia
Symptom	失眠 insomnia, 头痛 headache, 恶心 nausea, 咳嗽 cough, 鼻炎 rhinitis, 消化不良 dyspepsia, 便秘 constipation, 喘息 breathlessness, 更年期 climacterium, 流鼻涕 running nose, 上呼吸道感染 infection of the upper respiratory tract, 停经 menolipsis, 发烧 have a fever, 呼吸困难 expiratory dyspnea, 打喷嚏 sneeze, 昏睡 lethargy, 睡眠不足 lack of sleep, 鼻塞 rhinobyon
Dns	疾病 disease, 呼吸道 respiratory tract, 皮肤 skin, 鼻子 nose, 呼吸系统 respiratory system, 发病率 morbidity, 咽喉 throat, 心脏 heart, 疼痛 pain, 血液 blood, 喉咙 throat, 头部 head, 心血管 cardiovascular, 病人 patient, 病原体 pathogen, 病毒 virus, 耳鼻喉 ENT, 血液循环 blood circulation, 怀孕 pregnancy
Dv	生病 get sick, 刺鼻 nasal irritation, 伤肺 lung damage, 止咳 relieve a cough, 润喉wet whistle, 致癌 cause cancer

Seven types of diseases can be extracted from Table 3:

1) **respiratory tract disease**: 支气管炎 bronchitis, 咳嗽 cough, 喘息 gasps, 上呼吸道感染 infection of the upper respiratory tract, 呼吸困难 expiratory dyspnea, 呼吸道 respiratory tract, 止咳 relieve a cough;

2) **ENT disease**: 鼻炎 rhinitis, 打喷嚏 sneeze, 鼻塞 rhinobyon, 刺鼻 nasal irritation, 润喉 wet one's whistle, 鼻子 nose, 耳鼻喉 ENT, 咽喉 throat, 流鼻涕 running nose, 喉咙 throat;

3) **cancer**: 癌症 cancer, 肺癌 lung cancer, 致癌 cause cancer;

4) **pulmonary disease**: 肺癌 lung cancer, 肺炎 pneumonia, 伤肺 lung damage;

5) **skin disease**: 皮肤 skin;

6) **headache**: 头痛 headache, 头部 head;

7) **others**: 失眠 insomnia, 恶心 nausea, 消化不良 dyspepsia, 便秘 constipation, 停经 menolipsis, 发烧 have a fever, 昏睡 lethargy, 睡眠不足 lack of sleep, 发病率 morbidity, 疼痛 pain, 血液 blood,
病人 patient, 病原体 pathogen, 病毒 virus, 血液循环 blood circulation, 怀孕 pregnancy,
生病 get sick

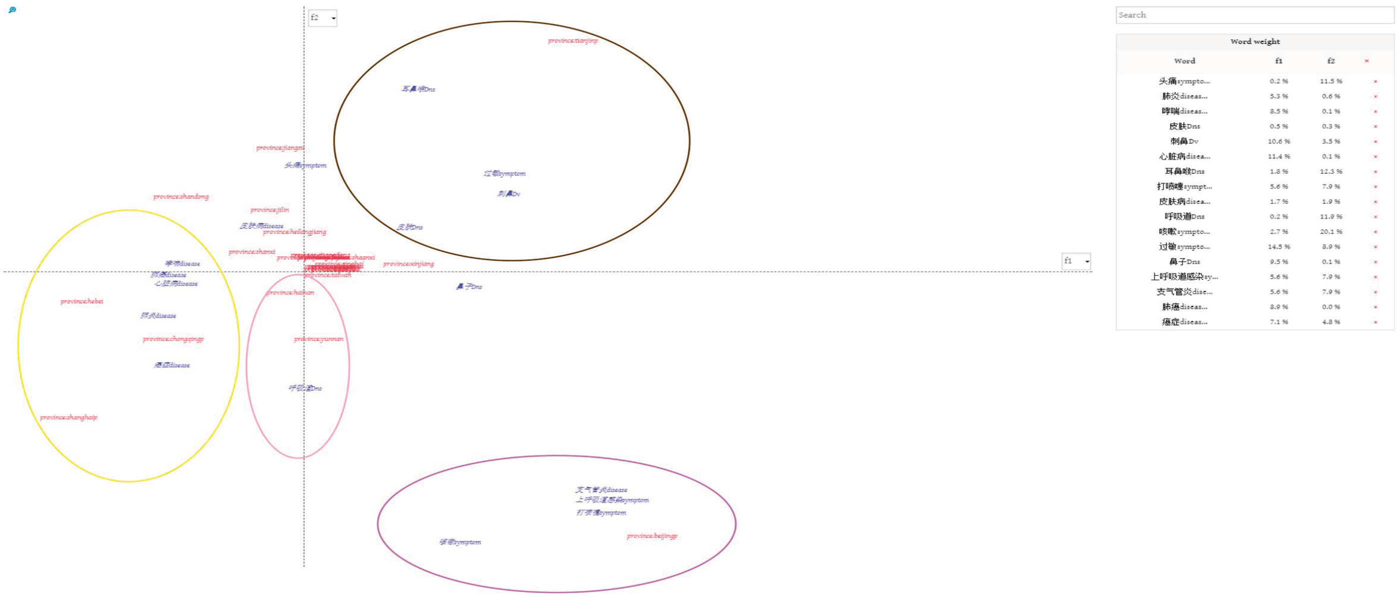
4. Methodology

According to the frequency of health issues related words, we have chosen five types of diseases (cancer, pulmonary disease, respiratory diseases, skin disease and cardiopathy), three sorts of symptoms (respiratory disease symptoms, allergy and headache) and 3 kinds of medical terminology (Ear Nose and Throat, skin, respiratory tract) with high-frequency in our corpus Weibo to study their distribution in 32 regions. In order to detect the correlation between region and health issues, we applied three HyperBase functions of (Salem 1982)[9]:

[9] HyperBase is a downloadable software for documentary and statistical exploration of texts. It was designed and developed by Étienne Brunet, assisted by Laurent Vanni. Created in 1989, HyperBase is mainly used for research of linguistics, literature, history, sociology or political science. We employed its online version, available at: http://hyperbase.unice.fr/hyperbase/doc/#/.

4.1 Analysis with the FCA model

FCA model is used to obtain a synthesized panoramic view of interactive attraction between regions and health issues. FCA means factorial correspondence analysis (Benzecri 1982), which provides a typology based on the different parts of text, allowing us to observe the lexical information in a subtler way, then it closes up the parts using the same words in the same proportions (Brunet 2011). The analysis results for a simultaneous representation of different lexical data are shown in a visual graph in the form of a scatter data plot (also called a scatter diagram) so that we can estimate the proximities calculated between different parts (Fig.2).



Search

Word weight			
Word	f1	f2	×
头痛sympto...	0.2 %	11.5 %	×
肺炎diseas...	5.3 %	0.6 %	×
哮喘diseas...	8.5 %	0.1 %	×
皮肤Dns	0.5 %	0.3 %	×
刺鼻Dv	10.6 %	3.5 %	×
心脏病disea...	11.4 %	0.1 %	×
耳鼻喉Dns	1.8 %	12.3 %	×
打喷嚏sympt...	5.6 %	7.9 %	×
皮肤病disea...	1.7 %	1.9 %	×
呼吸道Dns	0.2 %	11.9 %	×
咳嗽sympto...	2.7 %	20.1 %	×
过敏sympto...	14.5 %	8.9 %	×
鼻子Dns	9.5 %	0.1 %	×
上呼吸道感染sy...	5.6 %	7.9 %	×
支气管炎dise...	5.6 %	7.9 %	×
肺癌diseas...	8.9 %	0.0 %	×
癌症diseas...	7.1 %	4.8 %	×

Fig. 2. FCA of Weibo corpus – association of health issues and regions (Health issues are represented by blue words; Regions are indicated by red words).

With the analysis results of Weibo data revealed in the FCA graph (Figure 2), showing the approximate or distant relationships of different data points placed on the plan. In the framework of our analysis, this FCA graph is a matrix of specities of each health issue in each region. In order to show which heath issues are most associated with relevant regions, these two targets are plotted on the same axe. According to the positions of our analysing data, four 'regions+health issue terms' groups of interaction surrounded by different coloured circles:

Table 4: Groups of "provinces+health issues' combination

Group 1	Group 2	Group 3	Group 4
上海 shanghai 河北 hebei 重庆 chongqing + 哮喘 asthma 肺癌 lung cancer 心脏 cardiopathy 肺炎 pneumonia 癌症 cancer	海南 hainan 云南 yunnan + 呼吸道 respiratory tract	天津 tianjin + 耳鼻喉 ENT 过敏 allergy 刺鼻 nasal irritation	北京 beijing + 支气管炎 bronchit 上呼吸道感染 upper respiratory tract infection 打喷嚏 sneeze 咳嗽 cough

Through the FCA graph configured by figure 2, the interactive correlation between regions and health issues is polygamous, i.e. one or more regions are related to one or more lexicons of health issues. The FCA graph gives us a panoramic and combinatorial view of the correlative relationships between regions and health issues.

4.2 Analysis results with DISTRIBUTION

If we want to better understand the interactive attraction between each region and certain relevant health issues, we need to use DISTRIBUTION, a function built into HyperBase that allows us to visualize the distribution of health issue terms by regions. If the column presents a positive distribution in a section of the region, that means the attached word is overemployed in this part of the text, and vice versa (Fig.3).

We've chosen 17 health issues (癌症cancer, 肺癌lung cancer, 肺炎 pneumonia, 支气管炎 bronchit, 哮喘 athsma, 皮肤病 skin disease, 心脏病 cardiopathy, 咳嗽 cough, 上呼吸道感染 upper respiratory retract infection, 打喷嚏sneeze, 过敏 allergy, 头痛 headche, 耳鼻喉 ENT, 呼吸道 respiratory retract, 皮肤 skin, 鼻子 nose, 刺鼻 nasal irritation) to explore their distribution in 4 regions (Beijing, Shanghai, Hebei and Tianjin).

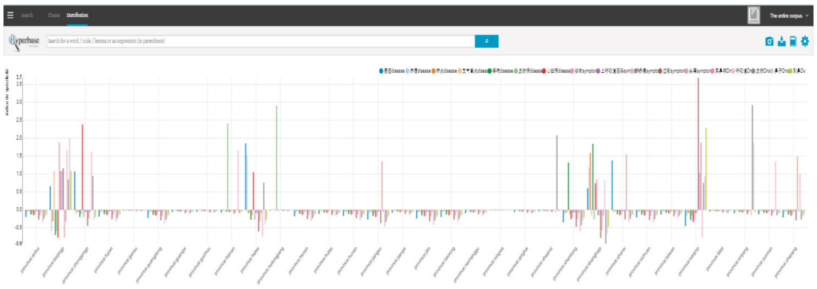


Fig. 3. Histogram of distribution of 17 health issue-related terms by regions.

Through the histogram of Figure 3, we found that health issue terms are distributed mainly in four areas: 北京 beijing, 上海 shanghai, 天津 tianjin and 河北 hebei. In addition, we noticed that these are also the four places with the highest frequency of occurrences according to our analysis results in Section 3. (Fig.1) Selection of investigative regions on geographic distribution of Weibo smog-related messages. Referring to our precedent analysis work on the FCA graph (Fig. 2), we can thus see in detail this histogram of health issue terms' distribution by regions (Fig. 3), and explore which health issue terms are specific or characteristic in each of those four most mentioned areas. When we compare different partitions of a corpus, rather than appealing to word frequency, we highlight the specific forms with their index of specificity since it is difficult to evaluate variable frequency of vocabularies, especially in consideration of the distinctive size of different parts of a corpus.

1) Analysis of 北京 beijing

As the capital of China, we have discovered that Beijing had the most numerous occurrence of health issue words: 癌症 cancer, 支气管炎 bronchitis, 咳嗽 cough, 上呼吸道感染upper respiratory tract infection, 打喷嚏 sneeze, 过敏 allergy, 呼吸道 respiratory tract, 皮肤 skin, 鼻子 nose, 刺鼻 nasal irritation. These 10 words comprise three types of health issues: cancer with 癌症 cancer; respiratory disease with 支气管炎 bronchitis, 咳嗽 cough, 上呼吸道感染 upper respiratory tract infection, 打喷嚏 sneeze, 呼吸道 respiratory tract, 鼻子nose, 刺鼻 nasal irritation; and skin disease with 皮肤 skin. Since the distribution all of these words are positive, we could therefore summarize that these three types of health issues are characteristic in Beijing.

2) Analysis of 上海 shanghai

Shanghai, a port city located in east of China, is the country's financial centre. Shanghai is among the most dynamic regions in China in terms of personal Gross Domestic Product (GDP). Through the result of statistical calculation, Shanghai is also plagued by air pollution with seven health issues shown in the histogram: 癌症 cancer, 肺癌 lung cancer, 肺炎 pneumonia, 哮喘 asthma, 心脏病 cardiopathy, 咳嗽 cough, 呼吸道 respiratory tract. Compared to the health problems exposed in Beijing, Shanghai contains not only cancer (癌症 cancer), respiratory disease (哮喘 asthma, 咳嗽 cough, 呼吸道 respiratory retract) as characteristic health issues related-terms, but also two types of diseases in addition: heart disease (心脏病 cardiopathy) and lung disease (肺癌 lung cancer, 肺炎 pneumonia).

3) Analysis of 河北 hebei

As a neighbour province of Beijing, Hebei's economic pillar is composed of coal and steel industry. With a monthly output of 19 million tonnes of steel[10] Hebei ranks in first place in the production of steel in China. Figure 3 shows four health problem terms exposed specifically in Hebei: 癌症 cancer, 肺癌 lung cancer, 心脏病 cardiopathy, 皮肤 skin. It should be noted that the only two terms about cancer issues both appeared in this province.

4) Analysis of 天津 tianjin

Located in the east of Hebei, Tianjin is also, like Beijing, directly exposed to air pollution coming from its neighbour Hebei in the west. In addition, Tianjin itself is an industrialized city. In Tianjin, we have 过敏 allergy, 头痛 headache, 耳鼻喉 ENT, 皮肤 skin, 鼻子 nose, 刺鼻 nasal irritation as specific terms of health problems, but two words – 过敏 allergy, 头痛 headache – appeared for the first time in the ventilation graph. In other words, compared to other areas of investigation, we have two new types of health problems – allergy and headache – showing in Tianjin.

4.3 Analysis results with WORDCLOUD

Thanks to WORDCLOUD (Fig.4), we could develop our research with some given health issues vocabulary to observe in a 'random' way their co-occurrence[11] (Vanni 2014), which means the co-occurring terms that are semantically approximate to a certain health-related word despite of its category. If in the mixed box of co-occurrences there are still one or several same areas co-occurring with the core word of health problems, just as in the interactive relationship seen in the distribution diagram in figure 3, a strong correlation between them will be confirmed by two statistical methodologies: Distribution of specific terms by regions and WORDCLOUD generated by co-occurrences.

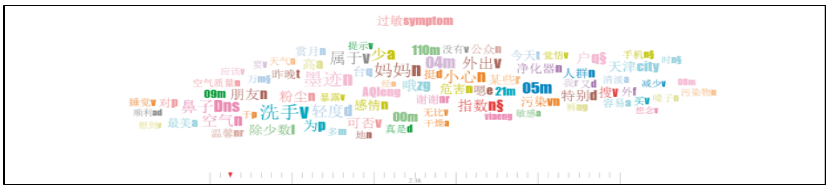


Fig. 4. Word cloud of co-occurrences of 过敏 allergy: 过敏 allergy + 天津 tanjin.

We've chosen the keyword of symptom 过敏 allergy to study his co-occurrences. Based on WORDCLOUD graphs generated by HyperBase, the correlative proximity between keyword and its co-occurrence depends on different sizes of the latter: the bigger the co-occurrence, the stronger the correlative relation indicated.

It should be noticed that the word cloud network constitutes a semantic field by associating the keyword and its co-occurrences. Thus, with the exception of the discovery of 'health issues+regions'(过敏 allergy co-occurs with 天津 tianjin) combinations, we also marked other lexical fields on the graph, for example, the word cloud of 过敏 allergy manifests four additional themes:

[10] Data Information from National Bureau of Statistics of China: http://data.stats.gov.cn/search.htm?s=%E6%B2%B3%E5%8C%97%E7%9C%81%E9%92%A2%E9%93%81%E4%BA%A7%E9%87%8F.

[11] Co-occurrence is the simultaneous presence of two or several words in the same statement (sentence, paragraph). It is used as an indicator of semantic proximity of these words. Co-occurrence networks provide a graphic visualization showing the potential relationships of collective co-occurrences of paired terms.

1) potential causes of allergy due to air pollution: 污染物 pollutants, 污染 pollution, 粉尘 ashes, 空气质量 air quality, 轻度 light (pollution), 空气 air, 干燥 dry, 指数 index (of pollution);
2) victim public: 公众 the public, 人群 the crowd, 妈妈 mother, 敏感 sensitive;
3) preventive measures adopted: 净化器 air purifier, 洗手 wash hands, 减少 reduce, 外出 go outside, 暴露 expose, 清淡 light (food).

5. Conclusion and discussion

In this article, with a data set of smog-related messages taken from the Chinese social network Sina Weibo, we sought to propose a conceptual and methodological framework that allows us to discover health issues exposed by Weibo, to investigate the geographic distribution of these health problems, and furthermore to explore the relations between smog-related health issues and their geographic distributions. For 18,898 smog-related Weibo from 64,098 words, extracted from the BCC database with seven search queries such as 雾霾 (smog), 大雾 (fog), 霾 (smog), 空气污染 (air pollution), 大气污染 (atmospheric pollution), PM2.5 (*Particular matter Ø 2.5* µm), PM10 (*Particular matter Ø 10* µm), our methodology – based on three complementary models (FCA, Distribution of specific words and WORDCLOUD of co-occurrences) – contributes to do a multidisciplinary work– lexical, discourse, geographic analysis – and to produce a set of cartography, histogram and word cloud graphs that demonstrate our observation and exploration. Finally, we obtained seven distinct types of relevant health issues, such as respiratory diseases, cancer, pulmonary diseases, ENT diseases, etc.； with respect to their geographic distribution, we have noticed that most of the health issues are intensively located in four regions – Beijing, Shanghai, Hebei, Tianjin – which play a strong part either due to their high frequency of occurrence in our corpus, or because of their political and economic position, or due to the reality of their air pollution conditions. As for the relationships between health issues and their geographic distribution, we also found that the three major health problems (pulmonary disease, respiratory disease and cancer) exposed by our Weibo data are distributed in the four predominant regions located in the north of China, which are plagued by atmospheric pollution. This study result proves that there exists an interactive correlation between certain types of health issues and their geographic location with regard to the smog problems in China.

References

Yang, L., Lin, Y., Lin, H. (2013). 基于情感分布的微博热点事件发 (*Micro-Blog Hot Events Detection Based on Emotion Distribution*).

Cui, A. (2013). 微博热点事件的公众情感分析研究 (*Study on Public Sentiment Analysis of Events in Microblogs*).

Shi, W., Wang, H. & He, Sh. (2012). 基于微博平台的公众情感分析 (Study on public sentiment based on microblog platform), *Journal of the China Society for Scientific and Technical Information*).

Wu, L.-Ch. (2014). L`opinion face à la crise Google dans les pays sinophones, *JADT*.

Signorini, A., Segre, A.M., Polgreen, P.M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS One*, *6*(5): e19467. Published online 2011 May 4. DOI: 10.1371/journal.pone.0019467.

Kim, E.K., Seok, J.H., Oh, J.S., Lee, H.W., Kim, K.H. (2011). Use of Hangeul Twitter to track and predict human influenza infection. *PLoS One*, *8*(7): e69305. Published online 2013 July 24. DOI: 10.1371/journal.pone.0069305. Print 2013.

Hồ-Đình, O., Valette, M. (2014). Analyse différentielle des discours de prévention du VIH : textes institutionnels et textes informels en français et en vietnamien, *JADT*.

Wang, S., Paul, M.J., Dredze, M. (2014). Exploring Health Topics in Chinese Social Media: An Analysis of Sina Weibo. *World Wide Web and Public Health Intelligence: Papers from the AAAI-14 Workshop*.

Xun, E., Rao, G., Xiao, X., Zang, J. (2016). 大数据背景下 BCC 语料库的研制 (T*he construction of BCC corpus in the age of Big Data*), 语料库语言学.

Salem, A. (1982). Analyse factorielle et lexicométrie. Synthèse de quelques expériences. *Mots*, № 4, 147-168.

Benzecri, J.-P. (1982). *L`nalyse des données/leçons sur l'analyse factorielle et la reconnaissance des formes et travaux*, Dunod.

Brunet, E. (2011). *Hyperbase – Manuel de Référence*, Université de Nice.

Vanni, L. et al. (2014). Arbres et co-occurrences. Nouvel outil logométrique sur le net. Application au discours de François Hollande. E. Néé, M. Valette, J.-M. Daube & S. Fleury (eds.), *JADT*, Paris, Inalco-Sorbonne nouvelle, 640-649.

Guerreau, A., Lafon, P. (1986).Dépouillements et statistiques en lexicométrie, *Histoire & Mesure*, Volume 1 – № 1, Varia, 105-110.

(Footnotes)

[1] Dns means special nouns in the category of Disease; Dv means verbs in the category of Disease.

Classical Arabic Authorship Attribution Using Simple Features

Saad Alanazi

College of Computer and Information Sciences, Jouf University, Saudi Arabia
Sanazi@ju.edu.sa

Abstract. The current paper aims to investigate the accuracy of a random forest approach to authorship attribution of Arabic poems. The dataset involves 100 poems by four poets from different countries: 25 poems for each. The language of all the poets is Classical Arabic. The 'Random Forest' is a trademarked process of averaging the results of multiple decision trees to improve the accuracy of the classification process. Twelve features were applied to the dataset. Although the overall accuracy of the data is satisfactory, with 76.4% for precision, 66.7% for recall, and 64.8% for F-measure, the results for precision and recall vary, from 54.5%-100%, and from 33.3-100%. The results suggest that the data are insufficient. It is also possible that the number of features extracted is insufficient. Future work should increase the data and features and evaluate the results against the findings of this study.

1. Introduction

The authorship attribution identifies the patterns of writing characteristic to a particular author (Juola, 2008). In theory, two writers will never compose in the exact same style, as authors will have different word preferences, spelling mistakes, levels of literacy, and grammatical accuracy, in addition to personal choices for punctuation. None of these differences are indicative on their own. However, these differences work together to shape an individual's writing style, which can be as unique as a fingerprint. Authorship attribution cares about identifying these defining characteristics and decoding such 'fingerprints', ensuring that authors can be accurately matched with the texts they have written.

Authorship attribution gains significance as digital texts are rapidly replacing written text. This means that traditional methods of identifying an author, such as by fingerprints, handwriting, or other physical proof, such as where the author was at the time the writing was penned, left, posted, or discovered, become less useful. Oftentimes, an author will be identified based on their Internet protocol (IP) address, email address, or phone number. However, on increasingly connected networks, where multiple users may access the same devices and where hacking is not rare, these addresses may not constitute strong evidence. Thus, we must resort to one measure to identify authors: their writing style. A person's writing style is very personal and allows us to attribute texts to the appropriate authors with relative accuracy. When the pool of possible authors is small, the accuracy of identification increases.

The oldest existing records of Classical Arabic literature come from poetry, with its special styles and structures dating back to the sixth century (Athamneh, 2017). Given that the Arabic traditions are primarily oral, only poetry has been preserved and memorised by people for its beauty and musicality. Such a literary advantage facilitates the memorisation process of the poetic lines, which was rampant among the narrators responsible for spreading the poems in the region.

Although Arabic poetry has been studied very closely since the eighth century, its complexity has always presented a challenge. There are two main types of Arabic poetry: rhymed or measured and prose (Al-Falahi et al., 2015). Rhymed poetry has 16 different metres, traditionally known as *buhuur* 'seas'. The units of seas are measured based on the number of *taf'ilas* 'metres', which must be included in every line, known as *bayt* (Al-

Falahi et al., 2015). This structure must be carefully respected, as any change in consonants or vowels could tamper with the poem, thus moving from one metre to another. In rhymed poetry, every line must end with the same rhyme. Scholar Al-Khalil Bin Ahmad Alfarahidi established formulae through which Arabic poetry is created and analysed (Al-Zahrani and Elshafei, 2012), yet these formulae are very complex, and many researchers aim to simplify them to make them easier to use. As for prose, it is characterised by its natural flow of literary speech and does not respect any rhythmic format.

Critics and researchers have not reached a consensus when it comes to the formulation, analysis, and categorisation of Arabic poetry. However, they agree on dividing it into 'classical' and 'modern' poetry. Classical Arabic poetry follows the traditional style and structures outlined by Al-Khalil Bin Ahmad Alfarahidi, regardless of which era it was written in. Modern Arabic poetry is a poetic version that is both recent (dating from the nineteenth to the twentieth centuries) and deviant from classical poetry structure into more prosodic flow influenced by the modernism movement.

In this study, I focus on classical rhymed poetry only, leaving the modern prosodic poetry for future work.

2. Related Work

Abbasi and Chen (2005) from the University of Arizona have observed the usage of Internet communication methods by extremists and the importance of proper authorship identification in these cases. They note that proper analysis can identify writing styles that are characteristic of terrorist communication. They also noted the importance of multilingual content analysis, as terrorist acts and therefore terrorist communication are international affairs.

Gazzah and Ben Amara (2008) investigated author attribution from hand-written Arabic texts, using a genetic algorithm to eliminate irrelevant features. They focus on authors' personal styles. Evaluating support vector machines (SVMs) versus multilayer perception (MLP) on 120 samples, they found that MLP was the strongest at 94% accurate.

Estival (2009) tested a prototype tool for author attribution on English and Arabic email corpora. Using a machine-learning method, she developed the classifiers. This tool was applied to identify the writing styles by demographics. The highest score came from measuring the education level, at 93.6% accurate. The lowest was personality traits, averaging 45% to 50% accurate.

Shaker and Corne (2010) proposed a set of Arabic function words for use in author attribution. Until then, function words were already used in English, but Arabic ones were created on a case-by-case basis. They selected 65 Arabic function words based on 104 English words. This allowed for more accurate Arabic author attribution.

Ouamour and Sayoud (2012) applied authorship attribution to ancient Arabic texts. They compiled a challenging dataset including authors from similar eras and demographics. Employing character n-grams and word n-grams as input for a sequential minimal optimisation (SMO) based SVM gave an 80% precision rating. They also found that they were able to apply similar rules to old Arabic as to modern English, highlighting that this method may be applicable to multilingual authorship attribution exercises.

Ouamour and Sayoud (2013) later ran some more tests on the same dataset. They used n-grams and rare words to identify authorship. They also applied seven classifiers, including Camberra distance, cosine distance, linear regression, Manhattan distance, MLP, SMO-based SVM, and Stamatatos distance. The SMO-SVM was 80% accurate, and MLP was 70% accurate. The worst were Camberra and Stamatatos, at 20% accurate. The rest were 60% accurate. This shows SMO-SVM classifiers are useful when applied to limited data.

Eltibi (2013) proposed an enhanced language model based on the probabilistic context free language model, adding more information on syntax and lexis. They also included a scoring function, measuring the importance of each feature. This allowed them to focus on individual author's styles over general style choices. This is very important, as a common style choices across authors would also be common for the authors themselves as individuals, which may confuse the data. Across 30 documents from nine authors, the model achieved 95% accuracy, which is 3.5% above the original model.

Otoom et al. (2014) also observed the application of intelligent software that identifies the author of a text. To solve the problem of limited research in Arabic author attribution, they proposed a new hybrid feature set. Following a few experiments for verification, they found that these sets have an 88% accuracy in the initial test and 82% accuracy in the cross-validation test.

Altheneyan and Menai (2014) observed the application of naïve Bayes classifiers to the field of authorship attribution. They were among the very first, if not the first, to apply naïve Bayes classifiers to authorship attribution in Arabic, investigating various methods. The models they used were naïve Bayes, multi-nominal naïve Bayes, multi-variant Bernoulli naïve Bayes, and multi-variant Poisson naïve Bayes. Applied to their own dataset, all naïve Bayes approaches displayed advantages over the existing methods. Multi-variant Bernoulli naïve Bayes had the strongest results, with 97.43% accuracy.

Al-Falahi et al. (2015) addressed the unique challenges of Arabic authorship attribution by running a classification task for Arabic poetry. The features observed included more common ones, such as characters, sentence length, and word length, but also less familiar ones, such as rhyme and the first word in a sentence. The training set involved a number of known authors, and the test dataset involved unknown authors, totalling 33 poets. The extra information allowed a precision rating of nearly 97%. Later, Al-Falahi and a reduced team (2016) achieved an even higher precision of nearly 99% by applying naïve Bayes, SVM, and SMO algorithms for 54 poets from different eras.

Omer and Oakes (2017) proposed a method based on an encoding of Arud, the metrical system of Arabic poetry, for the automatic authorship attribution. They have shown that the Arud-based encoding can be used as a linguistic feature for the automatic authorship attribution.

Observing the related research, we can see that many researchers have successfully applied a wide range of tools to the task of analysing Arabic literature. Of these, some of the most effective involved random forests, especially for analysing language that is not formal or literal. For example, Estival's 2014 application of the random forest approach to email written in Arabic. There is a gap in the research regarding poetry and other more artistic forms of literature. Therefore, the current study focuses on measuring the performance of the random forest methodology on Arabic poetry from different eras. Furthermore, as Classical Arabic poetry has a strict structure, this can complicate the process of authorship attribution, making it an interesting area to explore the full potential of the random forest approach applied to Arabic authorship attribution.

3. Methodology

Our methodology consists of four steps, which are data collection, data cleansing, features extraction, and classification. Figure 1 illustrates these steps.

3.1 Data Collection

The data were compiled from the Aldiwan website (www.aldiwan.net). Four poets were selected based on their large volume of work. A total of 100 poems were collected, 25 for each poet. This provides a wide variety of examples of Arabic poetry, which, due to the standardised structure of the genre, ensures that the author attribution task is sufficiently challenging. Table 1 below provides details about the corpus used in the experiment. The data were divided into two sets: a training dataset, which comprises 70% of the data, and a testing dataset, which constitutes 30%.

The chosen poets lived in the flourishing era of Arabic poetry. Their lives span from 756 to 1138 AD. Abu Nuwas lived from 756 to 814 AD and wrote most of his work during the reign of the Abbasside caliph Muhammad Al-Amin (809–813 AD). In his time, he was the best poet under the Islamic caliphate. Abu Firas al-Hamadani was the pen name of an Arabic prince, Al-Harith ibn Abi-l-Ala Sa'id ibn Hamdan al-Taghlibi, who lived from 932 until 968 AD. He held a high-ranking position in the court of Abbasside caliph Al-Muqtadir and was later an admired warrior and court poet. Ibn Zaydun lived from 1003 until 1071 AD. He lived across Cordoba and Seville and was considered the greatest neoclassical poet in the Arabic-owned territory of Al-Andalus in the Iberian Peninsula. He was born to an aristocratic family and was very involved in the political atmosphere. His poems document his own life and the political and social developments that took place during his lifetime.

Ibn Khafaja lived from 1058–1138, or possibly 1139. He is one of the most known poets of Al-Andalus and spent most of his life in the northern corner of the territory, near Valencia. Only when the Spanish occupied the territory around Valencia did he flee to North Africa. He lived a long life, and over his years of work, he developed some of the most sophisticated, original language that Arabic poetry had never seen until that time.

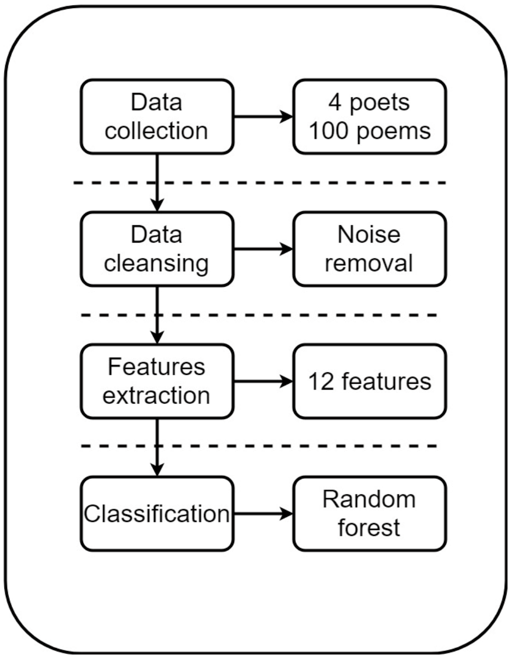


Fig. 1 Four-step methodology.

Although all four poets worked following the tradition of Classical Arabic poetry, which may confuse attempts at author identification, as the structure, number of verses, lines, and words, use of rhyme, etc., will be similar. However, they all come from different generations, led very different lives, and focused on different poetic traditions. This means that there will be differences in word use that may be detectable, allowing the right analysis to correctly attribute each poem to its author.

Table 1. Corpus details.

Poet	Number of poems	Number of verses	Number of words
Ibn Khafaja	25	626	5936
Ibn Zaydun	25	662	6309
Abu Firas	25	686	6036
Abu Nuwas	25	381	3537
Total	**100**	**2355**	**21818**

3.2 Data Cleansing

Until the ninth century, no written language made use of systematic punctuation. This means that documents that have been transcribed may have wildly difficult punctuation, depending on when the text was written and who transcribed it. Therefore, the punctuation has been eliminated in this study. This made the texts easier for the software to analyse and reduced the risk of false positives caused by differences in punctuation use.

3.3 Feature Extraction

It must be noted that, as we are dealing with Classical Arabic poetry, there is not a huge number of poems for each poet. Thus, using too many features would lead to overfitting given the size of the data used in the experiment. This restriction makes the number of features difficult to expand. Observing the literature, we can see that much larger numbers of features are typically used. For example, Estival (2014) used 518, and Altheneyan and Menai (2014) used 408. However, given the aforementioned reasons, 12 features were selected based on empirical experiments. They are as follows:

- *Metre*: in Arabic poetry, metre is determined not by syllables, but by counting the number of vowels and consonants in a consequent manner. The metre of Arabic poetry follows patterns strictly outlined by most writers during the years the four selected poets were composing their works. There were many different accepted metres that were considered acceptable, and different poets preferred different metres.
- *Rhyme*: In rhymed Arabic poetry, every verse must end with the same rhyme throughout the poem. There will be differences between rhymes preferred by different poets. There is also a set num-

ber of different rhyming schemes that were considered acceptable. Different poets have different preferences.

- *The most frequent letter:* Vowels are excluded because those sounds are the most frequent in any poem. Again, poets may have preferences based on individual taste, the subject the poet discusses, the era, and the region in which the poet lived. The website http://www.characterfrequencyanalyzer.com was used to calculate the frequency of the letters in each poem.
- *The second most frequent letter:* The second most frequent letter may vary depending on individual taste, the subject the poet discusses, the era, and the region in which the poet lived.
- *The third most frequent letter:* This may also vary depending on individual taste, the subject the poet discusses, the era, and the region in which the poet lived.
- *The part of speech (POS) tag of the first word in the poem:* Although metre and rhyme patterns are strongly controlled in traditional poetry, the words used in Arabic poetry are not restricted. This means that, in theory, a poem can begin with any word that is grammatically correct to start a sentence in Arabic. If poets prefer specific words or types of words, this will be a very obvious tool in authorship attribution.
- *The first letter in the poem:* This may vary depending on individual taste, culture, or the subject of the poem. It will be naturally related to the POS tag of the first word of the poem, especially if the first word itself is the same.
- *The number of verses:* The number of verses is not restricted in Arabic poetry. Despite this, there is a strong difference between the number of verses in Abu Nuwas's poems and those by the other three poets studied. It is possible that there are trends in the number of verses each author includes in their poems.
- *The length of the first word:* This may be more affected by the subject of the poem or the individual preference.
- *The average number of words per verse:* This is naturally going to be restricted by the strict standards that Arabic poems were held to. Because the number of letters and lines per verse are controlled, the average number of words per verse may be similar between different poets.
- *The average word length for each poem:* This will be much more varied. Depending on the poet's preference, the subject of the poem, the average word length, etc., will be different.
- *Grammatical state of the last word in each verse:* It is a special grammatical state of a noun, pronoun, verb, or adjective that reflects the grammatical function performed by that word in a phrase, or sentence. In Arabic, there are three main grammatical states, which are genitive, nominative, and accusative. Different poets had a preference for the grammatical state of the last word in each verse.

3.4 Classification

For classification purposes, random forest was applied using WEKA 3.8. A random forest consists of the averaging of multiple decision trees, each based on a random subset of unconnected variables.

The random forest approach was described by Breiman (2001). Since then, they have been applied to many fields due to their highly accurate predictions (Denisko and Hoffman, 2018). Random forest is an ensemble classifier made up of many decision trees, which averages the mode of the classes across these trees to output a class that applies to the group as a whole. The random forest approach was chosen because it makes the application of decision trees more precise. Individual trees have a habit of overfitting to their training set, whereas random forest adjusts to this by averaging data across the trees. This makes the random forest approach very fast to train (Friedman et al., 2001).

The data were split into training and testing sets: 70% and 30%, respectively. The Weka tool was used to perform the training and testing.

4. Results and Discussion

The results of the random forest, applied to author attribution, reach 76.4% for precision, 66.7% for recall, and 64.8% for the F-measure. The results are provided in Table 2. The results for precision and recall vary wildly, from 54.5% to 100%, and from 33.3% to 100%. This amount of variability shows that there was not enough data for a more accurate result. This would be the most likely reason for the variation of the results. The limited number of features would also lead to variable results. The results of our approach are illustrated in Figure 2.

Table 2. Results from random forests approach.

Poet	Precision	Recall	F-Measure
Ibn Khafaja	60%	75%	66.7%
Ibn Zaydun	100%	33.3%	50%
Abu Firas	54.5%	100%	70.6%
Abu Nuwas	83.3%	71.4%	76.9%
Total	**76.4%**	**66.7%**	**64.8%**

On the whole, these results show a lot of promise, as some of them are very strong. However, the inconsistencies show that more data and features could be required to improve accuracy with the random forest approach.

5. Conclusion

The results are promising, with an average of around 65% for the F-measure test. The general trend is strong for the precision, recall, and F-measure tests. However, the variability of the results illustrates that there are some limitations to the random forest approach carried out here. The main limitations are as follows.

The amount of data: Our dataset must be expanded to obtain a more accurate evaluation of the used method. It is highly likely that the wildly variable results are due to being limited to four poets with 25 poems each. Generally, increasing the training dataset would lead to a better prediction and classification.

The number of features: Our study is limited to only 12 different features. Surveying the literature reveals something interesting: increasing the number of features improves the strength of the trees. However, there is a trade-off between the accuracy and dimensionality/complexity, as each feature can only be acquired for an additional cost. To address the limitations for future work, the key change that needs to be made is increasing the size of the dataset used. It is possible that the number of features would also need to be increased.

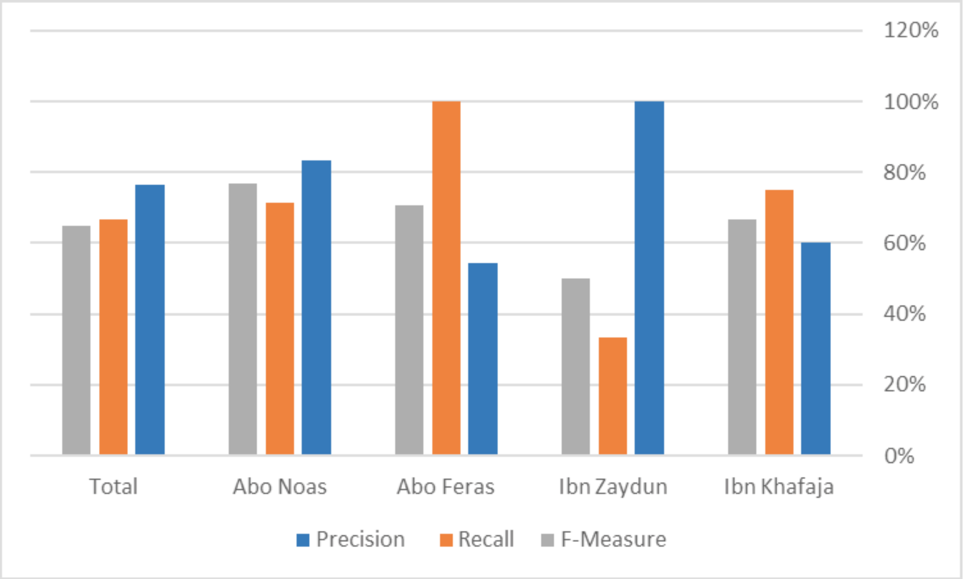


Fig. 2 Results of our approach.

References

Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, *20*(5), 67-75.

Al-Falahi, A., Mohamed, R., Mostafa, B., & Mohammed, A. (2015). Authorship attribution in Arabic poetry. *Intelligent Systems. Theories and Applications (SITA), 2015 10th International Conference*, 1-6.

Al-Falahi, A., Mohamed, R., & Mostafa, B. (2016). Authorship attribution in Arabic poetry using NB, SVM, SMO. *Intelligent Systems: Theories and Applications (SITA), 2016 11th International Conference*, 1-5.

Al-zahrani, S. & Elshafei, M. (2012). *Arabic poetry meter identification system and method*. U.S. Patent No. 8,219,386. Washington, DC: U.S. Patent and Trademark Office.

Altheneyan, A., & Menai, M. (2014). Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 473-484.

Athamneh, W. (2017). *Modern Arabic Poetry: Revolution and Conflict*. University of Notre Dame Press.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Denisko, D., & Hoffman, M. (2018). Classification and interaction in random forests. *Proceedings of the National Academy of Sciences*, *115*(8), 1690-1692.

Eltibi, M. F. (2013). *Author Attribution from Arabic Texts*. (master's thesis) Islamic University, Gaza, Palestine.

Estival, D. (2008). Author attribution with email messages. *Journal of Science*, Vietnam National University, 1, 1-9.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*, Volume 1, 337-387. New York: Springer series in statistics.

Gazzah, S. & Ben Amara, N. (2008). Neural networks and support vector machines classifiers for writer identification using Arabic script. *International Arab Journal of Information Technology (IAJIT)*, *5*(1).

Juola, P. (2008). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233-334.

Omer, A. & Oakes, M. (2017). Arud, the Metrical System of Arabic Poetry, as a Feature Set for Authorship Attribution. *In Computer Systems and Applications (AICCSA), 2017 IEEE/ACS 14th International Conference*, 431-436.

Otoom, A., Abdallah, E., Hammad, M., Bsoul, M., & Abdallah, A. (2014). An intelligent system for author attribution based on a hybrid feature set. *International Journal of Advanced Intelligence Paradigms*, 6(4), 328-345.

Ouamour, S., & Sayoud, H. (2012). Authorship attribution of ancient texts written by ten arabic travelers using a SMO-SVM classifier. *In Communications and Information Technology (ICCIT), 2012 International Conference*, 44-47.

Ouamour, S., & Sayoud, H. (2013). Authorship attribution of short historical arabic texts based on lexical features. *In Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2013 International Conference*, 144-147.

Shaker, K., & Corne, D. (2010). Authorship Attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. *In Computational Intelligence (UKCI)*, 2010 UK Workshop, 1-6.

A Shortest Path in an Experimentally Built Semantic Network

Maciej Godny[1], Izabela Gatkowska[1], Wiesław Lubaszewski[1]

[1] Jagiellonian University in Kraków, Poland
{m.godny, i.gatkowska, w.lubaszewski}@epi.uj.edu.pl
http://www.klk.uj.edu.pl/klk/en

Abstract. From linguistic perspective a semantic network built by the free word association experiment is a structure to explain a lexical meaning Deese (1965) and a human associating mechanism Clark (1970). It is interesting to find if network's path analysis may contribute to both hypotheses. Due to the nature of an experiment we restrict the analysis to shortest paths which can replace the stimulus – response connection, that is the only pairs produced by an experiment. We shall analyze a formal properties and a semantic consistency of the shortest path. We shall treat the network as an undirected weighted graph. All results are based on Polish experimental semantic network (Gatkowska, 2017).

1. Experimental Semantic Networks

One can observe the use of an experimentally built semantic networks in cognitive oriented NLP. For example network can be treated as a norm to evaluate a performance of algorithms which generate associations from text (Rapp, 2002; Wandmacher et al., 2008; Korzycki et al., 2017). One may also look into the network to find an information which is not lexically present in a text, e.g. *barks/barking* in lexical level implies a *dog* in semantic level (Lubaszewski et al., 2017). Finally, one may look at the network as a source to study human association mechanism.

Experimental semantic networks, e.g. English – Kiss (Kiss et al., 1973, Nelson et al., 1999), Dutch – DeDeyne and Storms (2008), Polish – Gatkowska (2015, 2017) are built by a use of the free word association test Kent-Rosanoff (1910), in which the tested person responds with a first word associated with a stimulus word provided by a researcher. The resulting network is a directed weighted graph. Each connection between two nodes has a direction, always from stimulus to response. Each connection has an association strength which is a number of participants, who are associated particular response with a particular stimulus, e.g. home – house 24 in EAT, i.e. first ever English experimentally built semantic network (Kiss et al., 1973).

Experimental semantic networks differ from manually built taxonomic networks as e.g. WordNet or CYC. Steyvers and Tenembaum (2005) had compared an experimental network Nelson et al. (1999) to WordNet and Roget's Thesaurus. The most important fragment of comparison which refers to an average number of connections for network node <*k*>, and average shortest path length *L*, is shown in Table 1.

Table 1. Comparison of experimental and taxonomic networks

	Nelson Undirected	Nelson Directed	Roget	WordNet
<k>	22.0	12.7	1.7	1.6
L	3.04	4.27	5.60	10.56

One may explain the difference in connection numbers <*k*> in following way. A taxonomic network uses mainly paradigmatic relations (subordinate – superordinate; part – whole) to link the nodes. Participant of the free word association experiment is not restricted and may use any type of semantic dependency to link the two words, which was observed by Deese (1965), Clark (1970). This answering freedom produces more connections per node in an experimental network, which means that description of a node meaning is richer, because we have multiple syntagmatic dependencies Clark (1970), which are properties like color, size, purpose etc. Gatkowska (2017). If we look at the *L* parameter we may say that the path length in taxonomic networks reflects a taxonomy depth, which does not exist in an experimental network.

2. Statistic Structure of an Experimental Semantic Network

The analysis of the formal properties of a experimentally built semantic network which was based on American English network (Nelson et al. 1999) was made by Steyvers and Tenembaum (2005). The comparative analysis of Dutch and English networks was made by DeDeyne and Storms (2008). Both analyses were performed on reduced networks, i.e. authors truncated all connections with association strength equal to 1, which means connections produced by a single person in the experiment. Both analyses considered a network as an unweighted graph. We shall use the criteria introduced by Steyvers and Tenembaum (2005) to analyze the Polish network (Gatkowska, 2017) and to compare Polish network to English and Dutch networks.

The criteria are as follows: n = the number of stimuli nodes; L = the average shortest path length, computed for each pair of all network nodes; D = the diameter of the network, i.e. the length of the longest path in all network shortest paths; <k> = the average number of incoming connections in directed network and in undirected network the <k> is the average number of connections to all neighbouring nodes. The Table 2. shows that compared networks are of different size. Nelson used 5018 stimuli, De Deyne 1424, Gatkowska 316 but the network structures seems to be similar.

Table 2. Network structure comparison.

	Nelson et al., 1999		DeDeyne, 2008		Gatkowska, 2017	
	directed	undirected	directed	undirected	directed	undirected
n	5018	5018	1424	1424	316	316
L	4.27	3.04	3.43	2.46	3.81	3.99
D	10	5	9	4	7	6
<k>	12.7	22.0	17	29	9.52	49.81

The differences in size of <*k*> are caused by differences in number of persons who responded to a particular stimulus i.e. 100–120 in Nelson experiment, 82–197 in DeDeyne and Storms, and 861–893 in Gatkowska.

3. The aim of the paper

Both papers cited above admitted that it would be interesting to analyze a shortest path in an experimental semantic network, but both did not analyze the shortest path in details. The aim of this paper is to present a formal properties and semantic consistency of the shortest path in an experimentally built semantic network. We shall make four assumptions.

First, to perform a shortest path analysis we have to refer to linguistic analyses of a network produced by the free word association experiment. Those analyses made from a linguistic point of view were based on assumption that each connection between stimulus and response is based on a semantic dependency between meaning of the stimulus word and the response word: Deese (1965), Clark (1970). This implies that a shortest path analysis may show unknown semantic properties of a network.

Second, we shall look only for shortest paths which may replace a stimulus – response direct connection. This kind of path may explain those stimulus – response connections that are not explicable by semantic relations, e.g. *baranina – wełna* (mutton – wool). This kind of path may enrich the meaning definition of the word that is a node of a network and may bring a data to study of human associating mechanism.

Third. To perform an analysis we shall treat a semantic network as an undirected graph. We know from linguistic analyses, e.g. Clark (1970), Gatkowska (2017) that direction of semantic relation between two words is independent from stimulus – response direction, i.e. if we have two pairs of different directions e.g. *bulldog*

→ *dog* and *dog* → *bulldog*, we can find that there is just one semantic relationship which goes from subordinate *bulldog* to superordinate *dog* (hyponymy).

Finally, we shall acknowledge that a semantic network built by the free word association experiment is a weighted graph, because each stimulus – response connection has an experimentally assigned association strength.

4. Experimental Semantic Network of Polish

Regarding network similarities shown in Table 2. Our analysis of a shortest path will base solely on Gatkowska (2017). In the experiment Gatkowska used 322 stimuli but we excluded from our analysis 6 expressions which were used as stimuli, i.e. *na świat, nie ma, za wsią, wuja Toma, do papieru, z dziurami*. We exluded also empty answers, created by participants who refused provide an answer to a particular stimuli. Finally, we excluded all answers with spelling error. Due to analyses made by Deese (1965), Clark (1970), Gatkowska (2015, 2017), which show that semantic quality of the stimulus – answer connection does not depend on association strength, we included all answers provided by a single person. The resulting network is described in Table 3.

Table 3. Polish network structure.

stimuli	nodes	connect dir	connect undir
316	12 182	41815	36 460

Due to nature of the free word association experiment in the undirected network the number of connections is equal to the number of outgoing connections which are present in directed network.

5. Shortest Path in an Experimental Semantic Network – Formal Properties

To perform an analysis we shall treat a semantic network as the undirected weighted graph. The shortest path between pair of nodes in the network means a path which starts with one node of the pair (start node) and ends with a second node of the pair (end node) and its path weight (the sum of weights of the path connections) is the smallest of all possible paths for a particular pair of nodes. For example in experimental semantic network for nodes *baranina – wełna* (mutton – wool) the path *baranina – baran – owca – puszysty – wełna* (mutton – ram – sheep – fluffy – wool) with path weight of 35 is shoter than path *baranina – mięso – lama – wełna* (mutton – meat – lama – wool) with path weight 39. If the two paths have the same path weight both are considered as a shortest path. In other words, it is possible in the semantic network that we can have more than one shortest path for a particular pair of nodes.

5.1 The Weight System

The association strength S is a number of test respondents who linked a particular response with the stimulus, for example if *dom – dach* (house – roof) were linked by 17 persons then S=17. One may convert an association strength into weight W using a total number of responses given to a particular stimulus T and it's relation to S. We shall use two methods of weight computing. The traditional one abbreviated as W1 (Kiss et al., 1973) where weight is computed as S/T; in our example of *dom – dach* where T = 872 and S = 17 the weight W1 = 0.019 (17/872). The second method abbreviated as W2 is computed as T/S., which means that for connection *dom – dach* W2 = 51.29 (872/17). In the case when two words are reversely a stimulus and a response, e.g. *kapusta* → *kiszona* (cabbage – saured) and *kiszona* → *kapusta* we have bidirectional connection with separate weights for each connection. In the undirected network we have to use only one weight. We decided to use a stronger one.

5.2 Path number

We shall use both weights (i.e. W1, W2) to find if the number of shortest paths which are in the network depends on weight. To test both weights we use three different values of association strength S between two nodes: S ≥ 1, S ≥ 2 and S ≥ 10. Figure 1. shows that path number in fact depends on S, because for both weights the number of paths differs substantially only with S ≥ 1.

If we consider fact, that in the experimental semantic network the weight of connection between the two nodes is set empirically, then we may add a new, more restrictive criterion in our analysis. Which means that we may look for the shortest paths which have the path weight smaller than the weight of direct connection between stimulus and response. Results of such restricted shortest path extraction are shown in Figure 2.

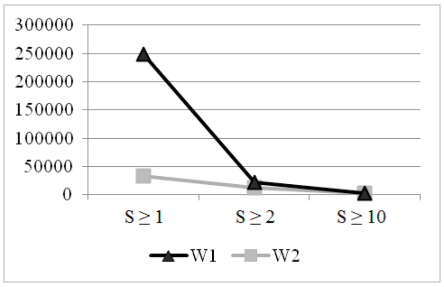


Figure 1. Path number.

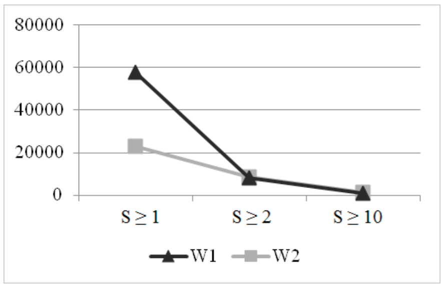


Figure 2. Path number – restricted.

As we can see the introduced restriction that relates path weight to the weight of direct connection makes the number of paths substantially smaller. We can also observe that both weights, i.e. W1 and W2 produce a similar paths number with S >= 2 and S ≥ 10.

To summarize one may say that for both weights W1, W2 there is a minimum value of S which produces a similar number of paths.

5.3 Path length

We shall treat a path length as a number of nodes in the path, including end and start nodes. We analyzed the path length for weights W1 and W2 with association strength S of values ≥ 1, 2 and 10. The relation between weight values and path length shows the Figure 3.

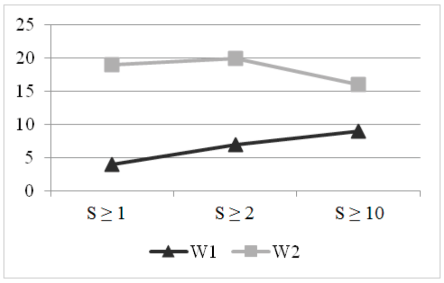


Figure 3. Maximum path length.

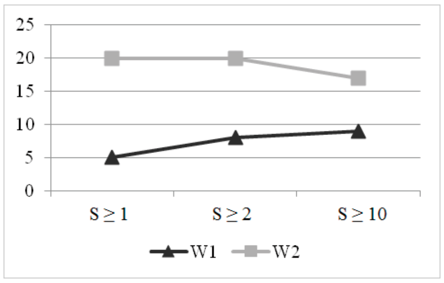


Figure 4. Maximum of the restricted path length.

Results show that the raise of weight value results in maximum of the path length. In the case of W2 we can observe that the maximum length of the path decreases. In the case of W1 we can see the opposite direction, the maximum length of the increases rapidly.

The detailed results are shown in Table 4, where pl means path length counted in nodes; weights W1 and W2 are counted as described above; S values are of ≥ 1, 2 and 10. Table shows a path number as a percentage of path total.

We can see in Table 4, that weight W1 produce much shorter paths than W2 – the longest path produced by W1 consists of 9 nodes compared to 20 for W2. For both W1 and W2 we can observe that the raise of the weight value results in increase of number of the short paths, i.e. of 3–4 nodes. But for W1 the number of paths longer that 4 also raises.

Table 4. Detailed unrestricted path length.

pl	S ≥	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
W1	1	16,7	83,3																
	2	22,9	48,1	22,4	5,6	0,3													
	10	31,5	35,1	15,3	11,3	4,3	2,0	0,5											
W2	1	15,8	17,4	10,8	13,2	11,2	10,1	8,2	5,2	3,5	2,0	1,2	0,7	0,4	0,2	0,1	0,1	0,1	
	2	23,1	21,9	10,6	11,8	8,9	8,3	6,1	3,8	2,5	1,3	0,8	0,5	0,2	0,2	0,1	0,1	0,1	0,1
	10	34,0	25,3	9,5	10,4	6,1	4,9	3,9	2,6	1,5	1,0	0,5	0,3	0,1	0,2				

Now we shall repeat path length analysis using restriction that says that path weight (the sum of weights of the path connections) should be smaller than the weight of the direct connection. Results are presented in Figure 4.

As we can see the condition does not change the observed earlier relation between the weight value and the path length. The detailed results are shown in Table 5. where path number is a percentage of path total.

Table 5. Detailed length of the restricted path.

pl	S ≥	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
W1	1	25,0	68,9	6,2															
	2	29,8	46,6	18,8	4,7	0,2	0,1												
	10	39,4	33,9	13,7	8,9	3,1	0,8	0,2											
W2	1	13,9	16,5	11,5	13,6	11,7	10,6	8,5	5,5	3,6	2,0	1,3	0,7	0,4	0,2	0,1	0,1	0,1	0,1
	2	21,6	21,7	11,7	12,0	9,2	8,7	6,3	3,7	2,4	1,2	0,8	0,5	0,2	0,1	0,1	0,1	0,1	0,1
	10	42,5	25,2	10,7	8,6	4,3	3,5	2,6	1,3	0,5	0,2	0,2	0,2	0,1	-	0,1			

Table 5. shows that the restricted extraction of the shortest path reduces path numbers (see Figure 2) but does not change relation between weight and path length.

5.4 Path Structure

5.4.1. Start Node and End Node structure

Due to lack of space we shall not provide an analysis of POS patterns which may occur in paths extracted from experimental semantic network. Where pattern means part of speech structure correlated with a path length. We shall restrict our analysis to the analysis of start and end nodes of the path regarding the three main parts of speech. We performed our analysis using path connection weights W1 and W2 with S ≥ 1. We show comparative results for both unrestricted and restricted shortest path extraction. Results are shown in Table 6. where N stands for Noun, V for Verb and A for Adjective.

Table 6. POS structure of the path.

	Unrestricted		restricted	
Sn- -En	W1	W2	W1	W2
N- -N	48%	48%	51%	51%
N- -V	2%	2%	3%	3%
N- -A	15%	15%	15%	15%
V- -N	4%	3%	4%	3%
A- -N	26%	26%	23%	23%
A- -A	3%	3%	3%	3%

First, looking at Table 6., we can find that the percentage of start node POS related to the end node POS does not depend on connection weight. One can suppose that the observed stability is an effect of stabile semantic dependency between the stimulus and response. Secondly, the observed results may support Deese's (1965) hypothesis that the set of responses provided by participants of the free word association experiment constitute the meaning of the stimulus word. And we know the word meaning is and should be stable.

5.4.2 Intermediate nodes in a path

In the case of intermediate nodes we are going to find all nodes which do not enter in direct connection with a start node in the network. In other words, nodes without a direct connection to a stimulus'. Such nodes are crossing the frontier of a subnet created by direct connections to a stimulus, which – according to Deese (1965) – define a stimulus meaning. We can mark automatically those intermediate nodes in a path. Results for both weights W1 and W2 for unrestricted path extraction are presented in table 7. Where we can see the number of paths which contain at least a single node which is not directly connected to start node are a large fraction of the total path number for both weights.

Table 7. Paths with intermediate nodes without a direct connection to a stimulus.

Weight	S ≥ 1	S ≥ 2	S ≥ 10
W1	70%	59%	45%
W2	66%	56%	57%

Having a global statistics of the paths extracted by an unrestricted algorithm we have to do an analysis of nodes which introduce those frontier crossing connections. We shall present a part of speech analysis – see table 8.

Table 8. A part of speech analysis.

	W1			W2		
	S ≥ 1	S ≥ 2	S ≥ 10	S ≥ 1	S ≥ 2	S ≥ 10
N	65%	73%	78%	83%	83%	83%
V	1%	2%	2%	1%	1%	1%
A	12%	20%	15%	15%	14%	15%
other	21%	5%	6%	1%	1%	1%

Table 8 shows that the frontier crossing connections are introduced mainly by nouns and other which are mainly proper names and pronouns. But such an observation does not explain how those frontier crossings relate to the meaning of a start node. In the next section we perform an analysis of the problem.

6. Semantic Consistency of a Path

As has been told above each pair of directly connected nodes is semantically connected due the nature of the free word association experiment which produces semantically related stimulus – response pairs. If we look at a path between two directly connected nodes where the start node is a stimulus and the end node is an answer we may find that some intermediate nodes in a path enter into semantic relationship with the start node and we may also find that some nodes are not related to the start node. If all intermediate nodes in a path are semantically related to the start node we may say that a path is semantically consistent, e.g. path *boleć – noga – ręka* (pain – leg – hand) where *leg* and *hand* enter the same 'location' relationship to *pain*. If there is an intermediate node in a path which does not enter in semantic relationship with the start node the path is not consistent, e.g. in the path *kapusta – zielona – roślina – kwiat* – roża* – czerwona,* (cabbage – green – plant – flower* – rose* – red) only two intermediate nodes enter in semantic relationship with the start node, i.e. *cabbage* 'color' *green* and *cabbage* 'hyponymy' *plant*. The inconsistent nodes are marked by asterisk (*). One may say that the inconsistent path in some point crosses the line which separates a subnet which defines a meaning of a start node as in Figure 5. where start and end nodes are darkened.

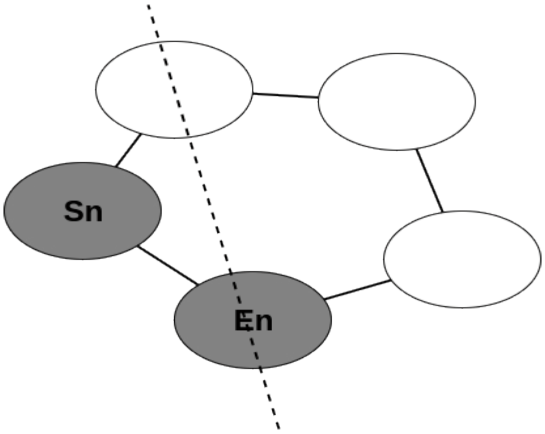


Figure 5. Subnet which defines a meaning of a start node.

One may also observe in our example of an inconsistent path that the intermediate nodes which do not enter into semantic relationship with the start node are not connected directly to the start node in the network. It would be interesting to find if those marked nodes caused that the shortest path extracting algorithm crossed the line which separates meanings in the network.

Unfortunately, analysis of path consistency should be done manually. We have no proper lexical tools to perform analysis automatically because WordNet uses only paradigmatic lexical relationships, i.e. hyponymy, meronymy and so on. On the other hand the FrameNet which describes syntagmatic dependencies does not use lexical relationships. Therefore we shall present results of manual analysis of shortest paths extracted by unrestricted algorithm for weight W2 with value of S ≥ 10, which means an analysis of 3260 paths. But regarding the similarities between W1 and W2 already shown one may think that this analysis is sufficient.

6.1 Consistency vs. directness of connection

Results of the path semantic consistency analysis are shown in table 9 where indirect stands for paths that contain at least one node which do not enter into direct connection with start node, and direct stands for paths with all nodes directly connected to a start node.

Table 9. Consistent vs. inconsistent paths.

	consistent	inconsistent
total	1765	1495
indirect	317	1161
direct	1448	334

First we can see in the table that semantically consistent paths share only 54% of extracted paths. Table 9 also shows that there are 18% of consistent paths which contain at least one node which do not enter into direct connection with a start node and there are 22% of inconsistent paths which are built with nodes directly connected to a start node. This implies that both direct and indirect connectivity do not fully explain path consistency. Finally, the fact that 82% of consistent paths were built from nodes directly connected to a start node rises doubts if those paths may expand Deese's definition of an associative meaning.

6.2 Path Consistency vs. Path Length

Table 10 presents results for 1765 semantically consistent paths evaluated manually, where column *Pl* (path length) presents node number in a path; columns next to the right show a number of indirect nodes in a path *In*, i.e. nodes which are not directly connected to the start node by a free word association experiment.

Table 10. Number of indirect nodes in semantically consistent paths.

Pl \ In	0	1	2	3	4	Total
3	973	0	0	0	0	973
4	381	180	0	0	0	561
5	61	46	16	0	0	123
6	32	36	21	1	0	90
7	1	3	4	4	0	12
8	0	2	1	1	**1**	5
9	0	1	0	0	0	1
	1448	**268**	**42**	**6**	**1**	**1765**

As we can see in the table 10 the path length is determined by a first node, which is indirectly connected to a start node (stimulus). Secondly, we may observe that the consistent paths which has a length greater than 5 share 6% of consistent paths total which means that there is no need for arbitrary fixed path length frequently used in analyses e.g. Steyvers and Tanembaum (2005).

Table 11. Number of indirect nodes in semantically inconsistent paths.

Pl \ In	0	1	2	3	4	...	Total
3	136	0	0	0	0	…	136
4	128	134	0	0	0	…	262
5	49	76	61	0	0	…	186
6	17	65	93	74	0	…	249
7	4	29	43	66	44	…	186
8	0	7	12	39	57	...	155
9	0	3	2	11	32	...	126
...	...	...	...	...	...	...	...
	334	**314**	**214**	**193**	**147**	**...**	1495

The table 11 shows that inconsistent paths are longer than consistent ones presented in table 10. We can also see that the path length is correlated with number of nodes indirectly connected to a start node. Finally, we can not say that path inconsistency is introduced by nodes indirectly connected to a start node, there are 334 (22%) without indirectly connected nodes. We think that a real reason of path inconsistency is a semantic value of direct connection between start node and end node, that is value of a connection between stimulus and response. The next section will bring a sketch of the problem.

6.3 Inconsistency Analysis

At this stage of the research we are not able to explain fully what makes a path a consistent or inconsistent. But we have found the following main situations in which the path may appear semantically inconsistent.

- a path is always inconsistent, when start node and end node are constituents of an idiom or multipart word, e.g. *czarny – owca* (black – sheep) which constitute the idiom *black sheep*. For this example we obtained paths like: *czarny – lęk – owca* (black– fear – sheep) or *czarny – kot – zwierzę – baran – owca* (black – cat – animal – ram – sheep):

- a path is frequently inconsistent when start node and end node enters in relation hyponymy (subordinate → superordinate), e.g. *owca – zwierzę* (sheep – animal) which produces paths like: *owca – owoc – zwierzę* (sheep – fruit – animal) or *owca – lew – zwierzę* (sheep – lion – animal).
- a path is frequently inconsistent, when start node and end node enter in relation meronymy (part – whole), e.g. *dach – dom* (roof – house), which produces paths like: *dom – szyba – dach* (house – glass – roof) or *dom – dym – dach* (house – smoke – roof).
- a path is frequently inconsistent, when the start node or end node is a feature that belongs to many objects, e.g. *duży – dom* (big – house) *duży – młot – dom* (big– hammer – house) or *duży – kozioł – dom* (big – goat – house).

Finally, we have to stress that our analysis is made to signal rather than solve the problem of the path semantic consistency.

7. Conclusion

The analyses presented in the paper are preliminary ones but they show that a pure statistical approach may not explain all properties of an experimental semantic network fully. But even such preliminary analysis suggests that only small fraction (18%) of consistent paths may bring an extension to a word meaning defined in the network. This implies also that a consistent path analysis may bring valuable data to study a semantic driven human association mechanism.

It seems to be clear that there is a need for further investigation of the semantic consistency of a path in the network to find if there is possible to create an algorithm which can extract only consistent paths in a network.

References

Clark, H.H. (1970). Word associations and linguistic theory. J. Lyons (ed.), *New Horizons in Linguistics*, Harmondsworth: Penguin Books, Harmondsworth.

DeDeyne, S., Storms, G. (2008). Word associations: network and semantic properties, *Behavior Research Methods*, *40*(1), 213-231.

Deese, J. (1965). *The structure of Associations in Language and Thought*, Baltimore, John Hopkins University Press.

Gatkowska, I. (2017). *Eksperymentalna sieć leksykalna języka polskiego* (*An Experimental Lexical Network of the Polish Languag*e), Kraków, Wydawnictwo Uniwersytetu Jagiellońskiego.

Gatkowska, I. (2015). Empiryczna sieć powiązań leksykalnych, *Polonica*, Volume 35, 155-178.

Kent, G. & Rosanoff, A.J. (1910). *A study of association in insanity. American Journal of Insanity*, *67*(2), 317-390.

Kiss, G.R., Armstrong, C., Milroy, R. et al. (1973). An Associative thesaurus of English and its computer analysis. A.J. Aitken, R.W. Bailey, N. Hamilton-Smith (eds.), *The computer and Literary Studies*, Edinburgh: University Press.

Korzycki, M., Gatkowska, I., Lubaszewski, W. (2017). Can the Human Association Norm Evaluate Machine-Made Association Lists? B. Sharp, F. Sèdes & W. Lubaszewski (eds.), *Cognitive Approach to Natural Language Processing*, 21-38.

Lubaszewski, W., Gatkowska, I. & Haręza M. (2015). Human association network and text collection. B. Sharp & R. Delmonte (eds.), *Proceedings of the 12th International Workshop on Natural Language Processing and Cognitive Science 2015*, Berlin, De Gruyter, 101-114.

Lubaszewski, W., Gatkowska, I., Godny, M., (2017). How a Word of a Text Selects the Related Words in a Human Association Network. B. Sharp, F. Sèdes & W. Lubaszewski (eds.), *Cognitive Approach to Natural Language Processing*, 41-62.

Nelson, D.L., McEvoy, C.L., & Schreiber, T.A. (1999). *The University of South Florida word association norms.* Retrieved from http://w3.usf.edu/FreeAssociation. Access date: September 25, 2017.

Rapp, R. (2002). The computation of words associations: comparing syntagmatic and paradigmatic approaches. *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, 1-7, Taipei, Taiwan, August 24-September 1.

Wandmacher, T., Ovchinnikova, E. & Alexandrov, T. (2008). Does latent semantic analysis reflect human associations? *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics*, Hamburg, Germany, August 4-8.

Steyvers, M. & Tenembaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, Volume 29, 41–78.

Biblioteka Jagiellońska
eISBN 978-83-949716-8-7