LREC 2018 Workshop

Linguistic and Neuro-Cognitive Resources (LiNCR)

PROCEEDINGS

Edited by

Barry Devereux, Ekaterina Shutova, Chu-Ren Huang

ISBN: 979-10-95546-08-5
EAN: 9791095546085

8 May 2018

Proceedings of the LREC 2018 Workshop
"Linguistic and Neuro-Cognitive Resources (LiNCR)"

8 May 2018 – Miyazaki, Japan

Edited by Barry Devereux, Ekaterina Shutova, & Chu-Ren Huang.

http://lincr2018.cbs.polyu.edu.hk/LiNCR_workshop/

Organising Committee

- Chu-Ren Huang (The Hong Kong Polytechnic University, Chair)
- Christophe Pallier (INSERM-CEA, CNRS, Co-Chair)
- Alessandro Lenci (University of Pisa)
- Barry Devereux (Queen's University Belfast)
- Brian Murphy (Queen's University Belfast)
- Chia-ying Lee (Academia Sinica)
- Ekaterina Shutova (University of Cambridge)
- Enrico Santus (SUTD-MIT)
- Jie-Li Tsai (National Chengchi University)
- John Hale (Cornell/DeepMind)
- Kathleen Ahrens (The Hong Kong Polytechnic University)
- Laurent Prévot (LPL-CNRS, AMU)
- Leila Wehbe (UC Berkeley)
- Mark Liberman (LDC, University of Pennsylvania)
- Philippe Blache (LPL-CNRS, AMU)
- Qin Lu (The Hong Kong Polytechnic University)
- Reinhold Kliegl (Potsdam University)
- Shi-Kai Hsieh (National Taiwan University)

Programme Committee

All Organizing Committee Members, plus the following:

- Ekaterina Shutova (University of Cambridge , Chair)
- Barry Devereux (Queen's University Belfast, Co-Chair)
- Adam Pease (Articulate Software)
- Alex Huth (UC Berkeley)
- Anna Korhonen (University of Cambridge)
- Chetwyn Chan (The Hong Kong Polytechnic University)
- Chia-Lin Lee (National Taiwan University)
- Chia-ying Lee (Academia Sinica)
- Emmanuele Cherisoni (Aix-Marseille University)
- I-Hsuan Chen (The Hong Kong Polytechnic University)
- Karl Neergaard (LPL-CNRS, Aix-Marseille University)
- Lorraine K. Tyler (University of Cambridge)
- Luana Bulat (University of Cambridge)
- Marco Senaldi (Normale Superiore di Pisa)
- Massimo Poesio (Queen Mary University of London)
- Natalia Klyueva (The Hong Kong Polytechnic University)
- Noël Nguyen (LPL-CNRS, AMU)
- Patricia Lichtenstein (UC Merced)
- Renkui Hou (The Hong Kong Polytechnic University)
- Shichang Wang (Shandong University)
- Stefan Frank (Radboud University)
- Vesna Gamez-Djokic (UC Berkeley)
- William S.Y. Wang (The Hong Kong Polytechnic University)

- Yao Yao (The Hong Kong Polytechnic University)
- Yunfei Long (The Hong Kong Polytechnic University)
- Yu-Yin Hsu (The Hong Kong Polytechnic University)

Preface

The goal of the LiNCR (pronounced 'linker') workshop is to provide a venue to share and explore a new generation of language resources which link and aggregate cognitive, behavioural, neuroimaging and linguistic data. Research activities within the purview of the workshop include the development of methods for the integration of neuro-cognitive data on language function with linguistic facts, the interpretation of experimental data when linked to rich linguistic information, and demonstrations of how new insights can be drawn from this powerful approach in domains such as language learning and neuro-cognitive deficits.

The papers and abstracts accepted for presentation in this workshop all showcase the potential of this interdisciplinary and data-driven framework for understanding language. A variety of complementary approaches are considered: building quantitative knowledge representations for linguistic phenomena using behavioral rating data; linking fMRI language data and computational linguistic measures; temporal alignment of linguistic and behavioral data; combined EEG and linguistic data on idioms and syntactic violations; ontological approaches to linking word meaning and human sensory experience, and investigating eye-tracking data as a tool for evaluating distributional semantic models and investigating lexical and syntactic processing.

We anticipate that our workshop will appeal to linguists who are interested in how the neurobiology of speech and language can constrain and inform theoretical accounts of linguistic phenomena, as well as cognitive neuroscientists and psycholinguistics who are interested in utilizing modern computational modelling and large-scale quantitative linguistic data in their experiments and data analysis pipelines.

As part of the programme, the ISI-NLP2 and LiNCR Workshops are collaborating to jointly support an invited talk and panel. Please refer to LiNCR/ISI-NLP2 website and proceedings for more details. An additional joint panel is also being planned together with the B&R LRE workshop.

C. Huang, C. Pallier, E. Shutova, B. Devereux, A. Lenci, B. Murphy, C. Lee, E. Santus, J. Tsai, J. Hale, K. Ahrens, L. Prévot, L. Wehbe, M. Liberman, P. Blache, Q. Lu, R. Kliegl, S. Hsieh May 2018

Programme

First Session

09.00 – 09.10 Opening Remarks

09.10 – 09.30 Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache
Event Knowledge in Sentence Processing: a New Dataset for the Evaluation of Argument Typicality

09.30 – 09.50 Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, Annabel Nijhof, Roel Willems
The Narrative Brain Dataset (NBD), an fMRI Dataset for the Study of Natural Language Processing in the Brain

09.50 – 10.10 Thierry Chaminade, Laurent Prévot, Magalie Ochs, Birgit Rauchbauer, Noël Nguyen
Challenges in Linking Physiological Measures and Linguistic Productions in Conversations

10.10 – 10.30 Philippe Blache, Stephane Rauzy, Deirdre Bolger, Chotiga Pattamadilok, Sophie Dufour
A Dataset for Studying Idiom Processing with EEG

10:30 – 11:00 *Coffee break*

Second Session

11.00 – 11.20 Adam Pease and Chu-Ren Huang
Ontology and Synesthesia: Language, Sense and the Conceptual Inventory

11.20 – 11.40 Amir Bakarov
Can Eye Movement Data Be Used As Ground Truth For Word Embeddings Evaluation?

11.40 – 12.05 Cory Shain, Marten van Schijndel, William Schuler
Deep Syntactic Annotations for Broad-Coverage Psycholinguistic Modeling

12.00 – 12.20 Charmhun Jo
Synesthetic Metaphors in Korean Compound Words

12:20 – 12:25 Chun-hsien Hsu, Chia-ying Lee, Jie-li Tsai
Frequency and Predictability Effects in Natural Reading: Evidence from Co-registration of Eye-movement and Event-related Potentials Measures

12:25 – 12:30 Zude Zhu, Xiaopu Hou, Yiming Yang
Reduced Syntactic Processing Efficiency in Older Adults in Reading Sentences

12:30 – 12:35 Libo Geng, Lillian Zhao, Jiaoyan Fang
The Stroop-like Effect During Sound Perception Task in Bilingual Minds

12:35 – 12:45 Concluding remarks and general discussion

The ISI-NLP2 and LiNCR Workshops are collaborating to jointly support an invited talk and panel. Please refer to LiNCR and ISI-NLP2 websites for more details.

Table of Contents

Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality

Paolo Vassallo*, Emmanuele Chersoni†, Enrico Santusˆ, Alessandro Lenci*, Philippe Blache†
University of Pisa*, Aix-Marseille University†, Massachusetts Institute of Technologyˆ
paolovassa@virgilio.it, emmanuelechersoni@gmail.com
esantus@mit.edu, alessandro.lenci@unipi.it, blache@lpl-aix.fr

Abstract

In the NLP literature, the *thematic fit estimation* task is defined as the task in which a system has to predict how likely a candidate argument (e.g. *cop*) is to fit a given a verb-specific role (e.g. the agent of *to arrest*) (Santus et al., 2017).
Because of the scarcity of benchmark datasets, thematic fit models are currently evaluated by measuring the correlation between their output and human ratings for isolated verb-filler pairs (Sayeed et al., 2016). However, such evaluation does not account for the dynamic nature of argument expectations: there is robust psycholinguistic evidence that human update their predictions on upcoming arguments during sentence processing, depending on the way other verb arguments are filled (Bicknell et al., 2010; Matsuki et al., 2011). Consider, for example, how the expectation for the patient of *to check* would change if we use *journalist* or *mechanic* as agents.
In this paper we introduce DTFit (Dynamic Thematic Fit), a dataset of human ratings for verb-role fillers in a given event context, with the aim of providing a rigorous benchmark for context-sensitive argument typicality modeling. The dataset accounts for the plausibility of patient, instrument and location roles, given the agent and the predicate.

Keywords: thematic fit modeling, distributional semantics, argument expectations, computational psycholinguistics, sentence processing, linguistic resources

1. Introduction

The psycholinguistic literature of the last two decades has brought extensive evidence for the cognitive relevance of the notion of *thematic fit*, that is to say the degree to which a given lemma fits in a given verb-specific role. A number of studies reported behavioral effects proving that, during on-line sentence processing, hearing a verb induces human subjects to activate expectations about nouns typically filling its thematic roles, and argument nouns in turn activate expectations about their typical predicates and typical co-arguments (McRae et al., 1998; Ferretti et al., 2001; McRae et al., 2005; McRae and Matsuki, 2009; Hare et al., 2009). These findings have been explained by researchers in the light of a *Generalized Event Knowledge* contained in the human semantic memory, which includes information about events and their participants (see Figure 1 for a summary of the priming effects). Such knowledge is activated by lexical cues in the sentences, and it is exploited by human subjects to anticipate the upcoming linguistic input (McRae and Matsuki, 2009).

More recent studies by Bicknell et al. (2010) and Matsuki et al. (2011) showed that verb argument expectations depend on the way other arguments are filled, and they are dynamically updated while the sentence is processed. For example, given the verb *to check*, if *journalist* is the filler of the agent role, then we can expect *spelling* or *report* to be very likely patient fillers. For the same verb, if the agent is *mechanic*, the most likely fillers will be things such as *brakes* and *engine*. Bicknell et al. (2010) presented a self-paced reading and an Event Related Potential (ERP) experiment where they compared sentence pairs differing only for their agents: their results show that sentences with a typicality relation between the agent and the patient are read faster by human and evoke smaller N400

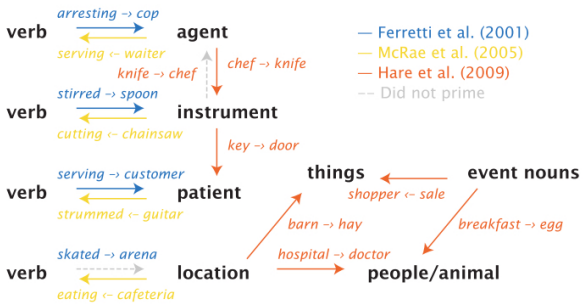

Figure 1: Summary of the experiments on event-based priming, from McRae and Matsuki (2009). The arrows between verb and roles indicate the direction of priming, from the prime to the target.

components[1]. Matsuki et al. (2011) set up a similar experiment with the self-paced reading and the eye-tracking paradigm, but focusing on the typicality relation between instruments and patients (e.g. *She used the shampoo to wash her hair* vs. *She used the shampoo to wash her car*). Coherently, they found significantly faster reading times for patient nouns that were more predictable given a predicate-instrument pair. Moreover, they reported shorter first fixation and gaze duration times for the patient in the eye-tracking experiment.

[1]N400 components are negative ERP deflections that peak around 400 milliseconds after the presentation of a stimulus word. The amplitude of the component elicited by a word has been found to be in an inverse relationship with its cloze probability (Kutas and Hillyard, 1984), and thus it has been considered as an index of the difficulty of integrating the meaning of a word in a given semantic context (Baggio et al., 2012).

The phenomenon of the thematic fit has recently raised interest also in the NLP community. Several studies have developed methods for the automatic estimation of the compatibility between candidate arguments and verb roles, generally adopting an evaluation based on the correlation between system predictions and human ratings. However, most of such work did not take into account the dynamic aspect of the phenomenon, i.e. the fact that the plausibility of arguments changes as the other roles are filled. The main reason behind such limitation is that the current gold standards mostly consist of simple ratings of verb-argument pairs in isolation, and do not take into account how the typicality scores change in function of the other event participants. In the present contribution, we precisely aim to address this issue by introducing the **DTFit dataset** (Dynamic Thematic Fit), a resource that has been built by specifically asking human subjects to produce plausible fillers for verb roles. Similarly to the previous literature, we collect data for the following roles: *patient, instrument* and *locations*, given the agent and the predicate. Our event tuples describe typical and atypical events, differing by just one argument (either the patient, the instrument or the location), and they are associated with human judgements collected in a Crowdflower task. Currently, we are still expanding the dataset, as we started the collection of judgements for new sets of tuples including new instruments and locations. Another planned expansion will regard the *time* role.

The paper is organized as follows. First, we illustrate the methodology and the criteria that guided the data collection, providing some statistical information about the dataset. Then, we will describe two approaches for the evaluation of thematic fit models on DTFit, showing the usefulness of our dataset.

2. Related Work

The thematic fit task, in the last decade, has been typically addressed by means of Distributional Semantic Models (DSMs). To the best of our knowledge, Erk et al. (2010) were the first authors to introduce the evaluation of a thematic fit model in terms of the correlation with human-elicited ratings. The authors used a syntax-based DSM to compute the plausibility of each verb role-filler pair as the similarity between a candidate filler and previously attested fillers for the same role. Finally, they measured the correlation of the system scores with a gold standard consisting of the human judgments collected by McRae et al. (1998) and Padó (2007).

One of the most influential frameworks for thematic fit modeling was the Distributional Memory by Baroni and Lenci (2010) (DM), which is also based on a syntax-based DSM. In the approach adopted by the authors, for each verb-specific role a *prototype vector* is built by averaging the syntax-based vectors of the most typical role fillers. The higher the cosine similarity of a noun with a role prototype, the higher its plausibility as a filler for that role.

Despite its simplicity, the method by Baroni and Lenci (2010) proved to be extremely effective, and inspired several extensions. Sayeed et al. (2015), for example, tried to improve the prototype representation by using vector features based on semantic roles, instead of syntactic dependencies. Moreover, they were the first to test to evaluate the plausibility of the fillers for roles other than the agent and the patient one, by introducing in the literature the Ferretti datasets for instruments and locations (Ferretti et al., 2001). Some other works aimed at addressing the problem of verb polysemy, either by obtaining different prototypes for the different senses through the hierarchical clustering of the fillers (Greenberg et al., 2015), or by testing similarity metrics based on a weighted feature overlap between the dimensions of the vectors (Santus et al., 2017).

It should be pointed out that all the above-mentioned works compare the scores of their systems with human rating for role-filler pairs in isolation: for example, given the patient role of the verb *to cut*, the rating quantify how good is *meat* as a filler. But as we anticipated above, the fitness of a filler depends also on the general event context: if we knew that the agent in the *cut*-event is a *government*, we would probably expect patients like the *taxes*, the *spending*, the *aids* etc. This aspect of dynamic update of the expectation on the fillers, at the present state, has received relatively little attention in the literature.

One of the few proposals addressing the dynamic update was given by Lenci (2011), who extended the original DM model (Baroni and Lenci, 2010) to account for the composition and update of argument expectations. Lenci tested an additive and a multiplicative model of vector composition (Mitchell and Lapata, 2010) to model the agent-related change in the expectations on the patient filler, by using a dataset derived from the sentences of the Bicknell experiment (Bicknell et al., 2010). The Bicknell sentences were turned into subject-verb-object triplets, such as *journalist-check-spelling*, and a binary classification task was set up for the evaluation. More concretely, the system had to measure the plausibility of a patient for pairs of triplets that differ only for the agent noun. It is important to notice that all triplets presented plausible patients with respect to the given predicates: only the agent made the patients of the respective triples more or less plausible. Therefore, the triples in each pair were found either in the *typical* or in the *atypical* condition, as in the following example:

(1) a. *journalist-check-spelling* (typical)
b. *mechanic-check-spelling* (atypical)

The goal, for each pair, was to identify the triple describing the most typical situation and in the end a global accuracy score was computed for each model.

Another system that was tested on the task of the argument expectation was the neural network architecture by Tilk et al. (2016), which was trained to generate probability distributions over selectional preferences for each thematic role. The authors used the Bicknell dataset as a benchmark, obtaining performances comparable to the multiplicative model by Lenci (2011). The same dataset was finally used by Chersoni et al. (2017), who have implemented some variations of Lenci (2011)'s system to demonstrate that DSMs benefit from structural information (i.e. syntactic information, to be intended as opposite of bag-of-words and bag-of-arguments hypotheses) when composing and updating thematic fit expectations.

To sum up, only a few studies so far have addressed the

problem of dynamic argument expectations, and the Bicknell triplets are currently the only available standard for testing their models.

3. The DTFit Dataset: The Data Collection Procedure

The only benchmark for the task of the argument expectation update, the Bicknell dataset, is limited in the sense that it allows evaluation only in terms of binary choice, i.e. it tests systems just on the capability of recognizing which argument combinations out of two is more typical, and it includes only agent and patient fillers. On the other hand, traditional thematic fit datasets include a wider variety of roles and more fillers for each role, also allowing researchers to perform an evaluation in terms of correlation (i.e., typicality is conceived as a score in a continuum rather than as a binary choice). However, such datasets only consist of verb-specific role-filler pairs and do not take the event context into account. Ideally, the DTFit dataset should combine the qualities of both resources.

In the sentence processing literature, several findings related to argument typicality have been shown to involve both aspects: the update of the expectation based on the event context (the saturation of an role can make a potential filler of another role more or less likely) and the priming relations (see the summary in Figure 1) between the events and the fillers of a wide variety of roles (McRae and Matsuki, 2009; Bicknell et al., 2010; Matsuki et al., 2011; Paczynski and Kuperberg, 2012).

Since our resource has the the advantages of both evaluation strategies (the information on event typicality and a more complex event context on the one hand, human ratings on multiple fillers for a given role on the other hand), we believe it will a useful benchmark for linking distributional models of event knowledge and experimental results.

3.1. Agents and Patients

To start our data collection, we parsed the corpus of image descriptions introduced by (Young et al., 2014). We decided to use this corpus as we wanted to have human-generated descriptions of typical visual scenes (i.e. images taken from Flickr). Then we have extracted from the corpus a list of verb-patient pairs, and we have selected 329 pairs for which it seemed intuitive to imagine a typical agent for the given scenario. For each pair, we produced a typical agent. Then, we created a another set of triples by replacing the original patient of each triple, in order to obtain corresponding atypical combinations (examples in Table 1).

agent	verb	patient	condition
mason	build	house	typical
mason	build	snowman	atypical
cook	clean	fish	typical
cook	clean	window	atypical

Table 1: Examples of triples produced starting from the pairs *build house* and *clean fish*.

In a second phase, we set up two Crowdflower task to obtain typicality ratings both for our agent-verb pairs and for our triples. Collecting judgements for agents and verbs alone was necessary, of course, to check that the perceived typicality of the event was depending on the noun filling the patient role.

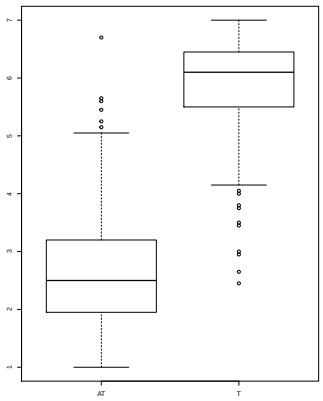

Figure 2: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Patients dataset.

As for the first task, we created two sets of 160 and 159 pairs, respectively. Each subset was rated by a group of 20 native speakers of British or American English. The subjects had to answer questions in the form *How common is for a **mason** to **build** something?*, by assigning a score between 1 (very uncommon, very atypical) and 7 (very common, very typical).

The second task was also taken by groups of 20 native speakers of British and American English. The triples were splitted in four subsets of 168, 168, 161 and 160 items, respectively, equally divided between typical and atypical ones. In this case, the questions had the form *How common is for a **mason** to **build** a **house**?*, and the subject had to provide an answer by using a seven-level Likert scale, as in the previous task.

As a check, we introduced 8 synonymy question for each test set, with the goal of filtering out the answers provided by trolls or non-attentive users. All the questions had the form *can **x** and **y** mean the same thing?* (e.g. *can "help" and "entertain" mean the same thing?*), and they were randomly presented to the subjects while taking the test. The responses of the subjects having less than a 70% accuracy in answering these questions were automatically excluded. With this strategy, we obtained typicality ratings for all our 657 triples, so we had to check the two conditions differ significantly. We compared the scores for typical and atypical conditions with the Wilcoxon rank sum test, and the test confirmed that the ratings for the former are significantly higher ($W = 106186.5, p < 2.2e-16$; see the boxplots in Figure 2).

3.2. Instruments and Locations

As we already mentioned in the Introduction, processing advantages were not found only for typical agent-patient combinations, but also for other roles, e.g. Matsuki et al.

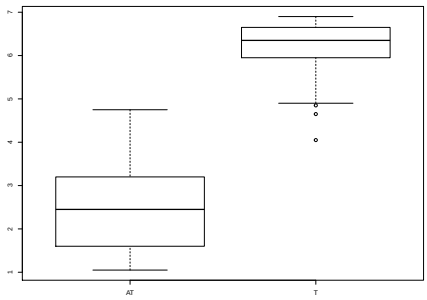

Figure 3: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Instruments dataset.

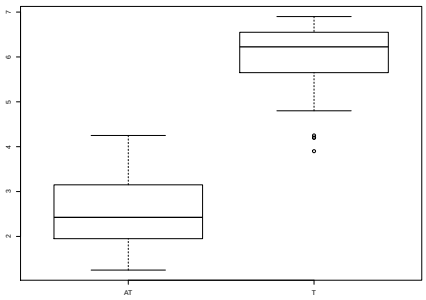

Figure 4: Comparison between the ratings for the atypical (AT, on the left) and the typical triples (T, on the right) in the Locations dataset.

(2011) found them also for sentences in which the patient was more predictable given a verb and an instrument.
From our dataset of agents, verbs and patients, we have thus selected two subsets of 50 triples, for which it was easy to imagine, respectively, typical *instruments* and typical *locations*. For each subset, we generated 100 quadruples by adding either a typical or an atypical argument to each triple (examples are shown in Table 2).

triple	**argument**	**role**	**condition**
mason mix cement	trowel	instrument	typical
mason mix cement	spoon	instrument	atypical
student drink beer	pub	location	typical
student drink beer	classroom	location	atypical

Table 2: Examples of quadruples produced by adding instruments and locations to the dataset triples *mason mix cement* and *student drink beer*.

We asked our subjects to rate the quadruples in the two dataset splits. The question, for each experimental item, was built according to the following pattern:

- how common is for a **agent** to use a **instrument** to **verb** a **patient**? (e.g. how common is for a *mason* to use a *trowel* to *mix cement*?)
- how common is for a **agent** to **verb** a **patient** in a **location**? (e.g. how common is for a *student* to *drink beer* in a *pub*?)

Also for these datasets, the test was taken by 20 native speakers of British or American English and synonymy questions were presented to the subjects as a check, as we previously described. The Wilcoxon rank sum test finally revealed significant differences between typical and atypical condition both for the Instruments ($W = 5, p < 2.2e-16$) and for the Locations dataset ($W = 6.5, p < 2.2e-16$).

3.3. Dataset Description

The current version of the dataset consists of three files:

- 656 triples of agents, verbs and patients (*Patients* dataset);
- 100 quadruples of agents, verbs, patients and locations (*Locations* dataset);
- 100 quadruples of agents, verbs, patients and instruments (*Instruments* dataset).

4. Evaluation Strategies

Previous studies on distributional models for thematic fit evaluated the systems either by measuring the correlation with human judgements or in a classification task. With our dataset, both evaluation strategies are possible:

- given a triple or a verb-filler pair, a system has to output a typicality/probability score and the performance will be evaluated as the correlation between the output scores and the human ratings that we collected in our Crowdflower tasks. Our resource allows to evaluate the typicality for verb-role fillers in isolation (e.g. the ratings for agent-verb and verb-patient combinations will be also made available in the future), but also to compose the expectations for an argument given a verb and the filler noun of another role (reflected by the rating of the entire triple);
- for each pair of triples sharing the agent and the verb, the system has to identify the triple with the higher score. Notice that this typicality has been shown to correspond to a processing advantage of the typical triples over the atypical ones, in terms of shorter reading times and of reduced amplitudes of the N400 component (Bicknell et al., 2010).

4.1. Baselines

In order to verify the quality of the dataset, we tested it by means of two DSMs derived from the approaches to thematic fit estimation by Lenci (2011) and Santus et al. (2017).
Both start by creating a prototype of the filler for a given verb-specific role by summing the distributional vectors of its typical fillers, updating the prototype on the basis of the information coming from the nouns saturating the other roles, and finally calculating the vector cosine with a candidate word. The latter adopts the same principle,

but it uses APSyn as similarity measure between the prototype and the candidate filler.[2] Both the systems were tested on two DSMs, namely the well known Distributional Memory (DM, Baroni and Lenci (2010)) and a dependency based DSM built from the co-occurrences extracted from the British National Corpus (Leech, 1992) and from the Wacky corpus (Baroni et al., 2009).

4.2. Experiments

DSMs. We have implemented two DSMs based on syntactic dependencies. One is based on the data of Distributional Memory (DM, Baroni and Lenci (2010)) and it includes co-occurrences between 30,490 target words and the same words in some syntactic relation with the target (since our space is just a slice of the original DM tensor, contexts have the form *dependency:word*). The other DSM was similarly built on the co-occurrences between 30,063 target words (we have selected the 30K most frequent nouns and verbs in our corpora, plus the words in the datasets) and the same words in some syntactic context.
This latter model, that we called **DEPS**, is a purely dependency-based model, in the sense that all the contexts have been automatically extracted as a syntactic co-occurrence between words. **DM**, on the other hand, has been enhanced with some manually-selected lexical patterns (e.g. *is-a, such-as* etc.).
TASKS AND EVALUATION. We measured the thematic fit for the three parts of our dataset:

1. the fitness of *Patients*, given the agents and the predicates (e.g. predict how likely is *toenail* as patient of *woman paint*);

2. the fitness of *Instruments*, given the agents, the predicates and the patients (e.g. predict how likely is *tray* as instrument of *waiter deliver drink*);

3. the fitness of *Locations*, given the agents, the predicates and the patients (e.g. predict how likely is *pub* as location of *student drink beer*).

The performances are evaluated in terms of both correlation of the scores and binary classification (i.e. typical tuples should get higher thematic fit scores than atypical ones, therefore we measure the accuracy of a system in assigning higher values to typical tuples). The first evaluation consists, concretely, in assessing the Spearman correlation between the scores delivered by our systems and the human ratings (see Section 3.1 and 3.2). The second evaluation consists in measuring the Accuracy of a each system in assigning a higher thematic fit score to typical tuples. This means that, for each dataset pair of tuples sharing the same verb and all the arguments but one, we score a hit each time the thematic fit score of the typical tuple is higher than the one of the corresponding atypical tuple (e.g. the score of *student drink beer pub* should be higher than *student drink beer classroom*).

[2] In their original paper, Santus and colleagues have also filtered the vectors according to certain syntactic relations, demonstrating that some relations contribute more than others to the identification of the similarity between the prototype and the candidate filler. We have however ignored this filtering step in our re-implementation.

Semantic Role	Syntactic Relation
Agent	Subject
Patient	Direct Object
Instrument	Complement introduced by *with*
Location	Complement introduced by *in,on,at*

Table 3: Summary of the syntactic relations that we used to select the typical role fillers.

PROTOTYPE. Following the method introduced by Baroni and Lenci (2010) and adapted by Santus et al. (2017), we measured the thematic fit as the similarity between the candidate filler and a prototype.
The prototype is either the sum or the multiplication between the sub-prototypes, which are vectors containing the sum of the distributional vectors of the most typical fillers for a role, given either the predicate or another argument. These role fillers are identified by means of a syntactic relation, which is used as an approximation of a deeper semantic role (the role-dependency mapping is summarized in Table 3): given a target word and a role, the k typical fillers are those with the highest PLMI association score (Evert, 2004) with the corresponding syntactic relation.[3] As in Baroni and Lenci (2010), we set $k = 20$ for all our models.
As an example, consider the computation of thematic fit for a triple like *mechanic check engine*:

- calculate the prototype of the most typical patient of *to check*, we select the 20 most typical objects of the verbs and sum their vectors;

- for the agent *mechanic*, we create another prototype by summing the vectors of the 20 most typical objects co-occurring with such an argument;

- the two prototypes are combined by either vector addition (**Add**) or vector pointwise multiplication (**Mult**);

- the resulting prototype is fed to the similarity measure, which calculates how similar it is to the candidate filler (in our case, *engine*).

In dataset 1. we have two "partial" prototypes to be combined, in 2. and 3. we have three of them. In other words, each additional argument introduces new information about the role to be predicted, and this information is encoded by means of a new prototype.
SIMILARITY MEASURES. The similarity measures adopted as thematic fit predictors are vector cosine, which is a standard metric for Distributional Models (Turney and Pantel, 2010), and APSyn Santus et al. (2017), which calculates the sum of the inverse of the average rank for each of the top N intersected features between two target vectors. As a value for this parameter, we present the results for $N = 2000$: this parameter value is a common choice in the previous literature and, also in this case, it gave the most stable performances across settings.

[3] Notice that the two models, **DM** and **DEPS**, use different labels to encode the relations, with different granularity.

DSM	Measure	Patients		Locations		Instruments	
		Add	Mult	Add	Mult	Add	Mult
DM	Cosine	**0.315**	0.29	**0.2**	0.17	**0.2**	0.11
	APSyn	0.27	0.29	0.17	0.13	0.127	0.146
DEPS	Cosine	0.287	0.22	0.17	0.12	**0.105**	0.05
	APSyn	0.33	**0.36**	0.13	**0.278**	0.04	0.09

Table 4: Spearman Correlation. In bold the best results by dataset and DSM; in bold and underlined the best scores by dataset.

Matrix	Measure	Patient		Location		Instrument	
		Add	Mult	Add	Mult	Add	Mult
DM	Cosine	67.97%	**69.28%**	61.22%	**63.26%**	**60%**	55.5%
	APSyn	65.03%	68.3%	57.14%	59.18%	55.5%	55.5%
DEPS	Cosine	68.67%	62.34%	54%	44%	**60%**	54%
	APSyn	66.77%	**73.4%**	62%	**66%**	50%	52%

Table 5: Accuracy in the binary classification task. In bold the best results by dataset and DSM; in bold and underlined the best scores by dataset.

4.3. Results and Analysis

Table 4 shows the results for the evaluation in terms of Spearman correlation. At a glance, it is clear that the correlation scores of our models in all settings are very low, proving that the task is a difficult one for DSMs. In particular, for the Instruments dataset no model achieve a correlation score above 0.2. This could be due to the fact that Instruments are often not expressed in event descriptions, and this could have led to the creation of more sparse prototype vectors for this dataset.

Concerning the performance, two models seem to perform more consistently: the additive models based on DM and vector cosine, and the multiplicative models based on DEPS and APSyn. The latter ones seem to take advantage from the multiplication operation, which sets to zero all the dimensions that are not shared by all sub-prototypes, and provides a similarity estimation based only on the dimensions that are "relevant" for all the other arguments.

As for the results for the binary classification task, they are shown in Table 5. Again, DM with cosine and addition and DEPS with APSyn and multiplication seem to perform more consistently than the others. Even in this case, the lowest performances overall are reported on the Instrument dataset, which confirms itself as the most difficult to model.

If we consider the two evaluation tasks together, it is clear that thematic fit estimation is a complex task for DSMs: for Instruments and Locations tuples, the correlation values with human judgements are extremely low and in the classification task no model manages to do significantly better than random guessing. [4] Future research on this topic might try to address the problem with more sophisticated approaches, i.e. neural network modeling.

5. Conclusion

In this contribution we have introduced DTFit, a new dataset for the evaluation of thematic fit estimation. The dataset has been designed having in mind the dynamic nature of the phenomenon, with the specific goal of providing a resource that allows for the evaluation of context-sensitive argument typicality. We used our dataset to test two different models, which has been shown in the previous literature to perform very well in the traditional evaluation settings for the thematic fit task. The results showed that our dataset is a challenging benchmark for classic syntax-based DSMs, and probably more sophisticated approaches will be required to improve modeling performances.

In the end, we are convinced that thematic fit modeling is an important task for bridging the gap between computational models and experimental results, and that the notion of distributional similarity can be used to model phenomena related to argument expectations (i.e. reduced reading times, or reduced N400 amplitudes for predictable arguments). We hope that our resource will turn out to be a useful tool for the research in computational psycholinguistics going in this direction.

[4] Verified with the Chi-Square test: for the best classifier, $p > 0.1$.

6. Acknowledgements

Emmanuele Chersoni's research is supported by the A*MIDEX grant (n. ANR-11-IDEX-0001-02) funded by the French Government "Investissements d'Avenir" program.

7. Bibliographical References

Baggio, G., Van Lambalgen, M., and Hagoort, P. (2012). The Processing Consequences of Compositionality. In *The Oxford Handbook of Compositionality*, pages 655–672. Oxford University Press.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, M., Bernardini, S., Ferraresi, A., and 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.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., and Kutas, M. (2010). Effects of Event Knowledge in Processing Verbal Arguments. *Journal of memory and language*, 63(4):489–505.

Chersoni, E., Santus, E., Blache, P., and Lenci, A. (2017). Is Structure Necessary for Modeling Argument Expectations in Distributional Semantics? In *Proceedings of IWCS*.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis.

Ferretti, T. R., McRae, K., and Hatherell, A. (2001). Integrating Verbs, Situation Schemas, and Thematic Role Concepts. *Journal of Memory and Language*, 44(4):516–547.

Greenberg, C., Sayeed, A. B., and Demberg, V. (2015). Improving Unsupervised Vector-space Thematic Fit Evaluation via Role-filler Prototype Clustering. In *Proceedings of HLT-NAACL*, pages 21–31.

Hare, M., Jones, M., Thomson, C., Kelly, S., and McRae, K. (2009). Activating Event Knowledge. *Cognition*, 111(2):151–167.

Kutas, M. and Hillyard, S. A. (1984). Brain Potentials during Reading Reflect Word Expectancy and Semantic Association. *Nature*, 307(5947):161.

Leech, G. N. (1992). 100 million words of english: the british national corpus (bnc).

Lenci, A. (2011). Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–66.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C., and McRae, K. (2011). Event-based Plausibility Immediately Influences On-line Language Comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(4):913.

McRae, K. and Matsuki, K. (2009). People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, 3(6):1417–1429.

McRae, K., Spivey-Knowlton, M. J., and Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, 38(3):283–312.

McRae, K., Hare, M., Elman, J. L., and Ferretti, T. (2005). A Basis for Generating Expectancies for Verbs from Nouns. *Memory & Cognition*, 33(7):1174–1184.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive science*, 34(8):1388–1429.

Paczynski, M. and Kuperberg, G. R. (2012). Multiple Influences of Semantic Memory on Sentence Processing: Distinct Effects of Semantic Relatedness on Violations of Real-world Event/state Knowledge and Animacy Selection Restrictions. *Journal of Memory and Language*, 67(4):426–448.

Padó, U. (2007). *The Integration of Syntax and Semantic Plausibility in a Wide-coverage Model of Human Sentence Processing*. Ph.D. thesis.

Santus, E., Chersoni, E., Lenci, A., and Blache, P. (2017). Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Sayeed, A., Demberg, V., and Shkadzko, P. (2015). An Exploration of Semantic Features in an Unsupervised Thematic Fit Evaluation Framework. *Italian Journal of Computational Linguistics*, 1(1).

Sayeed, A., Greenberg, C., and Demberg, V. (2016). Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of ACL Workshop on Evaluating Vector Space Representations for NLP*.

Tilk, O., Demberg, V., Sayeed, A., Klakow, D., and Thater, S. (2016). Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 37:141–188.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

The Narrative Brain Dataset (NBD), an fMRI Dataset for the Study of Natural Language Processing in the Brain

Alessandro Lopopolo[1], **Stefan L. Frank**[1],
Antal van den Bosch[1,2], **Annabel Nijhof**[3], **Roel M. Willems**[1,4,5]
[1]Center for Language Studies, Radboud University
Erasmusplein 1, 6525 HT Nijmegen, the Netherlands
[2]Meertens Institute, Royal Netherlands Academy of Arts and Sciences
Oudezijds Achterburgwal 185, 1012 DK Amsterdam, the Netherlands
[3]SGDP Centre, King's College London
16 De Crespigny Park, Denmark Hill, SE5 8AF London, UK
[4]Donders Institute, Radboud University
Kapittelweg 29, 6525 EN Nijmegen, The Netherlands
[5]Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, the Netherlands
{a.lopopolo, s.frank, a.vandenbosch}@let.ru.nl, annabel.nijhof@kcl.ac.uk, r.willems@donders.ru.nl

Abstract

We present the Narrative Brain Dataset, an fMRI dataset that was collected during spoken presentation of short excerpts of three stories in Dutch. Together with the brain imaging data, the dataset contains the written versions of the stimulation texts. The texts are accompanied with stochastic (perplexity and entropy) and semantic computational linguistic measures. The richness and unconstrained nature of the data allows the study of language processing in the brain in a more naturalistic setting than is common for fMRI studies. We hope that by making NBD available we serve the double purpose of providing useful neural data to researchers interested in natural language processing in the brain and to further stimulate data sharing in the field of neuroscience of language.

Keywords: fMRI, neuro-linguistics, naturalistic stimuli, narrative, perplexity, surprisal, PoS

1. Introduction

The Narrative Brain Dataset (NBD) is an fMRI dataset created by recording the brain activity of 24 native speakers of Dutch during passive listening to three narrative Dutch texts: excerpts from audiobooks. This task and these stimuli are intended to be as naturalistic as possible. The dataset is meant to be used by researchers interested in the study of natural language processing in the human brain using naturalistic, unconstrained linguistic material. This dataset has already been used in a number of neuroscientific studies combining computational linguistic models and brain imaging analysis techniques, as exemplified in Section 6 .
NBD comes with meta-data describing the temporal structure of the stimulus presentation (word onset, offset and duration) and with a series of supplementary annotation of the stimulus texts that might come useful as starting point for further analysis of the data.
We hope that by making NBD available we serve the double purpose of providing useful neural data to researchers interested in naturalistic language comprehension, and to further stimulate data sharing in the field of neuroscience of language.

2. Dataset Structure

The NBD dataset consists of three parts: fMRI data, text & meta-data, and supplementary annotation.
fMRI data (*/fMRI/*) contains 24 folders (*/S01/,..., /S24/*) – one for each subject. Each subject folder is divided in 6 run folders ((*/run1/,/run2/,/run3/,/run4/,/run5/, and /run6/*) containing .nii volume images constituting the magnetic resonance recording during the presentation of the stimuli. Table 1 explains the relation between runs and stimuli – in Section 3 2 we explain the procedure behind the 6 runs structure, whereas in Section 4 we give more details about the stimuli. The data is preprocessed according to the methods described in Section 3 and in a format that is compatible with SPM8 and later versions[1]. The current format can be easily converted into other formats according to the user's needs. We decided not to include the raw fMRI images for reasons of space and efficiency.

Run name	Stimulus	CGN name
run1	Narrative 1	fn1055
run2	Narrative 2	fn1100
run3	Narrative 3	fn1090
run4	Narrative 1 reverse	NA
run5	Narrative 2 reverse	NA
run6	Narrative 3 reverse	NA

Table 1: Correspondence between fMRI data runs, stimulus narratives (or reverse recordings of narratives) and original "Corpus Gesproken Nederlands" (CGN) file names.

Text & meta-data consists of three .txt tab separated files (*Narrative_1_wordtiming.csv, Narrative_2_wordtiming.csv, Narrative_3_wordtiming.csv*) containing the text of the three narratives presented to the subjects. Each row in the

[1] `http://www.fil.ion.ucl.ac.uk/spm/software/spm8/`

file corresponds to one word (or word + punctuation) of one of the three narrative text stimuli accompanied with the temporal parameters of each word of the textual stimuli with regard to the experimental paradigm described above. These consist of word onset, offset and duration in seconds. Table 2 provides an example of temporal meta-data for a sentence from the stimulus texts. These parameters are especially important given that a) fMRI volume acquisition and word onset are not synchronized; b) fMRI volume is acquired every 880 ms (see Section 3 3), whereas word duration is variable and can be shorter than that (as Table 2 exemplifies).

Word	Onset	Offset	Duration
plotseling	0.103	0.68	0.577
is	0.68	0.843	0.163
ze	0.843	0.976	0.133
er.	0.976	1.149	0.173

Table 2: Example of the timing information of the stimulus text.

NBD is also provided with a battery of **supplementary annotations** consisting of part of speech (PoS) tags and computational measures assigned to each word of the text stimuli (files: *Narrative_1_annotation.txt, Narrative_2_annotation.txt, Narrative_3_annotation.txt*). Each column of the annotation file contains: the word and its PoS, word frequency, PoS frequency, average phonetic frequency, word perplexity, PoS perplexity, average phonetic perplexity, word entropy, word semantic association. The procedures used to obtain these additional annotations are described in Section 5 .

3. Magnetic Resonance Data

3.1. Participants

Twenty-four healthy, native speakers of Dutch (8 males; mean age 22.9 years, range 18-31) without psychiatric or neurological problems, with normal or corrected-to-normal vision, and without hearing problems took part in the experiment. All participants except one were right-handed. Ethical approval was obtained from the CMO Committee on Research Involving Human Subjects, Arnhem-Nijmegen, The Netherlands (protocol number 2001/095), in line with the Declaration of Helsinki.

3.2. Procedure

The experimental paradigm consisted of passively listening to the three narratives (see Section 4) and their reversed versions (for a total of six sessions) inside the MRI scanner. That amounted to six experimental runs, all collected in one single fMRI session on the same day. Each story and its reversed speech counterpart were presented following each other. Reversed speech versions of the stories were created with Audacity 2.03[2]. Half the participants started with a non-reversed stimulus, and half with a reversed speech stimulus. Participants were instructed to listen to the materials attentively, which in practice is only possible for three narratives, and not for the reversed speech counterparts. There was a short break after each fragment.
Stimuli were presented with Presentation 16.2[3]. Auditory stimuli were presented through MR-compatible earphones. After the scanning session, participants were tested for their memory and comprehension of the stories.

3.3. Scanner Parameter

Images of blood-oxygenation level-dependent (BOLD) changes were acquired on a 3-T Siemens Magnetom Trio scanner (Erlangen, Germany) with a 32-channel head coil. Pillows and tape were used to minimize participants' head movement, and the earphones that were used for presenting the stories reduced scanner noise. Functional images were acquired using a fast T2-weighted 3D echo planar imaging sequence (Poser et al., 2010), with high temporal resolution (time to repetition: 880 ms, time to echo: 28 ms, flip angle: 14, voxel size: 3.5 $\times$ 3.5 $\times$ 3.5 mm, 36 slices). High resolution (1 $\times$ 1 $\times$ 1.25 mm) structural (anatomical) images were acquired using a T1 sequence.

3.4. Preprocessing

Preprocessing was performed using SPM8[4] and Matlab 2010b[5]. The first four volumes were removed to control for T1 equilibration effects. Rigid body registration was used to realign images. Images were realigned to the first image within each run. The mean of the motion-corrected images was then brought into the same space as the individual participant's anatomical scan. The anatomical and functional scans were spatially normalized to the standard MNI template, and functional images were re-sampled to 2 $\times$ 2 $\times$ 2 mm voxel sizes. Finally, an isotropic 8-mm full-width at half-maximum Gaussian kernel was used to spatially smooth the motion-corrected and normalized data.

4. Linguistic Data

Narrative text used as stimuli presented to the human subjects consisted of three excerpts from three distinct literary novels extracted from the Spoken Dutch Corpus, "Corpus Gesproken Nederlands" (CGN) (Oostdijk, 2000).[6]

The excerpts were spoken at a normal rate, in a quiet room, by female speakers (one speaker per story). Stimulus durations were: Narrative 1 (CGN file fn1005) 3:49 min, Narrative 2 (CGN file fn1100) 7:50 min, and Narrative 3 (CGN file fn1090) 7:48 min.

Table 3 contains summary information about the three narratives, including number of words, mean and range of word duration in milliseconds.

5. Annotation

Besides the temporal information, the linguistic data is accompanied by two additional types of annotation: linguis-

[2] http://www.audacityteam.org

[3] https://www.neurobs.com

[4] http://www.fil.ion.ucl.ac.uk/spm

[5] http://www.mathworks.nl

[6] Narrative 1: from Peper, R., *Dooi*, L.J. Veen, 1999; Narrative 2: from Van der Meer, V., *Eilandgasten*, Contact, 1999; Narrative 3: from Jakobsen, A., *De Stalker*, De Boekerij, 1999

		Word Duration (msec)	
	# Words	Mean (s.d.)	Range
Narrative 1	622	273 (181)	4-1174
Narrative 2	1291	252 (160)	31-949
Narrative 3	1131	274 (183)	40-1221

Table 3: Summary information of the three narrative texts used as stimuli.

tic – consisting of the PoS tags of the words in the text – and computational measures – consisting of stochastic and computational semantics measures computed on the word, PoS and phonological level of the texts.

5.1. Linguistic Annotation

The words in the stimuli are annotated with their syntactic categories, or parts of speech (PoS). The tagset employed here was the one employed by CGN (Oostdijk, 2000) and comprises 320 tag types[7]. Besides 13 base tags, this method explicitly assigns morpho-syntactic sub-category features to the base tags containing information such as gender, number, form and so on. This tagset closely follows the practices of the Dutch Grammar "Algemene Nederlandse Spraakkunst" (ANS) (Haeseryn et al., 1997).

5.2. Computational Annotation

All words in the linguistic data are assigned seven stochastic measures: word frequency and perplexity, PoS frequency and perplexity, average phonological frequency and perplexity, and word entropy. A measure of the semantic association between each word and its preceding textual context is also provided.

5.2.1. Stochastic Measures

Perplexity – the degree to which the actually perceived item x_t in a series deviates from expectation – is computed as an exponential transformation of the surprisal of encountering x_t given its previous context $x_1 \ldots, x_{t-1}$:

$$\text{ppl}(x_t) = 2^{\text{surprisal}(x_t)} = 2^{-\log P(x_t|x_1,\ldots,x_{t-1})}$$

The conditional probabilities required for obtaining perplexity are estimated by a second-order Markov model, also known as a trigram model. That is, $P(x_t|x_1, \ldots, x_{t-1})$ is simplified to $P(x_t|x_{t-2}, x_{t-1})$. Using SRILM (Stolcke, 2002), the model was trained on a random selection of 10 million sentences (comprising 197 million word tokens; 2.1 million types) from the Dutch Corpus of Web (NLCOW2012) (Schäfer and Bildhauer, 2012).
The PoS perplexity is computed analogously. Instead of using the surface forms of the training and stimulus set, the trigram model was trained on the PoS-tagged version of the same 10 million sentences subset of NLCOW2012. The tagging was performed using the Frog toolbox for natural language processing of Dutch text (Daelemans and van den Bosch, 2005; van den Bosch et al., 2007)[8].

Phonological perplexity was estimated from conditional probabilities $P(p_t|p_{t-1}, p_{t-2})$, where the ps refer to the phonological transcription of the words in the running texts into a sequence of phonemes using a memory-based grapheme phoneme converter (Busser et al., 1999) trained on CELEX 2 (Baayen et al., 1995). The probabilities are computed using WOPR[9] (van den Bosch and Berck, 2009) trained on CELEX 2 (Baayen et al., 1995). Once phoneme-wise perplexity is computed, the phonemic perplexity of each word of the stimulus is computed as the average value across the phonemes of that word.

Next-word **entropy** was also derived from the conditional probabilities of words given their preceding context. It is a function of the distribution of probabilities of all possible upcoming words. It is computed as:

$$E(x_{t+1}) = \\ -\sum_{x_{t+1} \in V} P(x_{t+1}|x_t, x_{t-1}) \log P(x_{t+1}|x_t, x_{t-1}),$$

where V denotes the vocabulary (i.e., the set of word types in the training data). Entropy values were computed by WOPR (van den Bosch and Berck, 2009).

5.2.2. Semantic Similarity Measures

The semantic similarity between each content word w_t and its preceding context C is computed as the cosine between the distributional semantic vector representations of w_t and of C. Semantic vector representations of words were generated by the word2vec skipgram model (Mikolov et al., 2013). The representation of C is defined as the sum of the vector representations of the four content words preceding w_t (or fewer, if w_t is among the first four words of the text). If w_t is the first content word of the text then C is empty so semantic distance is undefined.

6. Published Analyses of the Current Dataset

The present fMRI data has already been analysed in several studies, demonstrating that naturalistic linguistic tasks and fMRI can yield interesting and meaningful results. Willems et al. (2016) have shown that entropy and surprisal predict brain activity in different brain areas. Frank and Willems (2017) demonstrated that predictive measures (surprisal) and semantic association measures can be distinguished with regard to brain area sensitivity. Similarly, PoS, lexical and phonological stochastic measures divide the cortical language network in non-overlapping sub-networks (Lopopolo et al., 2017). Part of the data was used by Nijhof and Willems (2015) to investigate how individuals differently employ neural networks important for understanding others' beliefs and intentions, and for sensori-motor simulation while processing narrative language.

7. Data Availability

The NBD is available at `https://osf.io/utpdy/`.

[7] more details at `http://lands.let.ru.nl/cgn/doc_English/topics/version_1.0/annot/pos_tagging/info.htm`

[8] `http://languagemachines.github.io/frog/`

[9] `https://ilk.uvt.nl/wopr/`

8. Acknowledgements

The work presented here was funded by NWO Gravitation Grant 024.001.006 to the Language in Interaction Consortium and by a grant from the Netherlands Organisation for Scientific Research (NWO-Vidi 276-89-007).

9. Bibliographical References

Busser, B., Daelemans, W., and van den Bosch, A. (1999). Machine learning of word pronunciation: the case against abstraction. In *Sixth European Conference on Speech Communication and Technology, EUROSPEECH 1999, Budapest, Hungary, September 5-9, 1999*.

Frank, S. L. and Willems, R. M. (2017). Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Lopopolo, A., Frank, S. L., van den Bosch, A., and Willems, R. M. (2017). Using stochastic language models (slm) to map lexical, syntactic, and phonological information processing in the brain. *PLOS ONE*, 12(5):1–18, 05.

Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics, May.

Nijhof, A. D. and Willems, R. M. (2015). Simulating fiction: Individual differences in literature comprehension revealed with fmri. *PLOS ONE*, 10:1–17, 02.

Poser, B., Koopmans, P., Witzel, T., Wald, L., and Barth, M. (2010). Three dimensional echo-planar imaging at 7 tesla. *NeuroImage*, 51(1):261 – 266.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

10. Language Resource References

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

Daelemans, W. and van den Bosch, A. (2005). Memory-based learning in natural language processing. *Memory-Based Language Processing*, pages 3–14.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J., and van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst*. ONijhoff and Deurne: Wolters Plantyn.

Oostdijk, N. (2000). The spoken dutch corpus. overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May. European Language Resources Association (ELRA). ACL Anthology Identifier: L00-1083.

Schäfer, R. and Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, et al., editors, *LREC*, pages 486–493. European Language Resources Association (ELRA).

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. pages 901–904.

van den Bosch, A. and Berck, P. (2009). Memory-based machine translation and language modeling. *The Prague Bulletin of Mathematical LinguisticsE*, 91(17).

van den Bosch, A., Busser, B., Canisius, S., and Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In Frank V. Eynde, et al., editors, *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114. Leuven, Belgium.

Challenges in Linking Physiological Measures and Linguistic Productions in Conversations

Thierry Chaminade[1], **Laurent Prvot**[2,4], **Magalie Ochs**[3], **Birgit Rauchbauer**[1,2,5], **Nol Nguyen**[2]
[1]Aix Marseille Universit , CNRS, INT, Marseille, France
[2]Aix Marseille Universit , CNRS, LPL, Aix-en-Provence, France
[3]Aix Marseille Universit , CNRS, LIS, Marseille, France
[4]Institut Universitaire de France, Paris, France
[5]Aix Marseille Universit , CNRS, LNC, Marseille, France
firstname.lastname@univ-amu.fr

Abstract

We introduce here a new experimental set-up that provides temporally aligned linguistic and behavioral data together with physiological activity time-series recorded during social interactions. It brings the experimental approach closer to ecological social interaction. Such endeavour requires the aggregation of linguistic, physiological and neuro-cognitive information. Compared to measurement of activity grounded on existing linguistic material our setting presents some additional challenges as we are dealing with conversations. In addition to present the rationale, set-up and preliminary analyses, we discuss (i) the challenges caused by the spontaneous and interactional nature of the activity recorded ; (ii) the problem of balancing experimental set-up between the technical needs and the desire to keep some level of naturalness in the task ; and (iii) the difficulties in relating in a temporal way linguistic events with physiological signals that have their own biological dynamics.

Keywords: conversation, physiology, artificial agents

1. Introduction

We consider that *conversations* constitute a privileged framework to study social interactions. Our objective is to approach these highly sophisticated linguistic and social structures by scrutinizing neuro-physiologival responses of the participants as well as their observable behaviors such as gaze, facial expressions and verbal productions. Our ultimate goal is to characterize the participants' brain activity by means of fMRI but because of the huge challenges involved in using fMRI in conversational interactions, we started out with a simpler setting that featured some of the challenges, in particular in terms of data analysis. More precisely we recorded computer-mediated "skype like" conversations and we tracked a set of physiological parameters during these conversations : gaze (eye-tracking) and electrodermal activity. Electrodermal activity, as a physiological response, has a specific dynamic that needs to be handled while temporally relating the measures to the actual linguistic production in the corpus (Chaminade, 2017). In the study, we were interested in comparing different communication situation, in particular subjects were interacting either with another human or with an artificial agent. The significant differences we found between the two conditions is a interesting step showing the relevance of our measures and analyses.

2. Experimental set-up

A cover story provided a topic for the discussion as well as a common goal for the two interacting agents. The cover story consisted in presenting the experiment as a neuromarketing experiment, in which the pair of participants would discuss together through a videoconferencing system the message of a forthcoming advertising campaigns. In each campaign, the tested participant is presented with three images without text and then instructed to discuss it with either a natural (fellow human) or an artificial (embodied conversational agent or anthropomorphic robot) agent.

2.1. Physiological pilot set-up

Experimental conditions were defined by a 2 by 2 factorial plan. The first factor was the nature of the Agent the participant discussed with, a Human or an Artificial agent presented as autonomous; the second factor was the nature of the Interaction, either Live, through videoconferencing, or Video, using recordings of previous Live interactions as stimuli. The four conditions were therefore Human/Live, Artificial/Live, Human/Video, Artificial/Video. The virtual agent GRETA (Bevacqua et al., 2010; Pelachaud, 2015) used for the behavioral part of the project was used in a Wizard of Oz (WOZ) setting.

There was a room for the participant and another room for the human agent. Headphones were used so that the speech from both participants were acquired separately. In the Participant room, the recorded participant sat in front of a computer screen topped by the webcam and included microphone used for the Skype discussion. The Control computer was connected to the screen and the webcam and the participants headphones, which also controlled GRETA and WoZ. The eye-tracker cameras were below this screen. The left hand of the participant was fitted with a blood pulse sensor and two electrodes to record the electrodermal activity measurement guidelines (Roth et al., 2012), connected to Biograph box. The second room comprised a computer controlled by the Control computer, and was connected to the discussant screen, webcam and headphones.

2.2. Neurophysiological Experimental set-up

The participant is in the MRI scanner with earphones while the human agent is in another room with headphones, both

facing a screen. Several aspects of behavior and physiological responses of this participant are recorded as continuous time series to form the corpus. Speech production, eye movements and skin conductance of the scanned participant are recorded with MRI-compatible devices. Recording of eye movement offers, in addition to eye-tracking data set, a live video output of the eyes of the scanned participant for the interlocutor. The human interlocutor is recorded by a webcam with incorporated microphone. The artificial agent is the conversational robotic platform Furhat. 12 repetitions of the 60-second discussions with the human interlocutor and each version of the artificial agents are recorded for each participant using the fact that there are several images in the advertising campaign that must be discussed separately. Each repetition consists in the presentation of one image for 10 seconds to the participant, followed by a 3-second black screen, and then sixty seconds during which the participant talks with the interlocutor, either a human participant or the artificial agent controlled by a WOZ. Around 20 participants are used for the analysis of an fMRI experiment, yielding 12 minutes of discussion per subject per agent, for an expected total of 4 hours per agent.

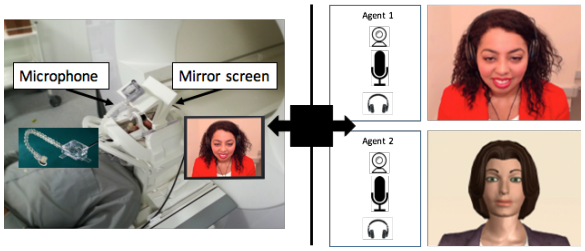

Figure 1: fMRI Experimental set-up

2.2.1. Equipment

fMRI requires dedicated MRI-compatible equipment: MR-compatible visual stimulation, eye-tracker, earphones and physiological response recorders (blood oxygenation and skin conductance, based on Siemens technology made available with the MRI scanner) (see Figure 1). An fMRI-compatible microphone with online de-noising has been acquired for completing the set-up, allowing for real-time discussion. The MRI center research ingeneer developed an interface to synchronize all this data. Preliminary analysis of the resulting speech recorded is very promising, with ASR system able to recognize significant parts of the input and no problem for manual transcription.

2.2.2. Remaining difficulties

There are two anticipated difficulties that call for specific attention. First, head movements interfere with analysis of the neurophysiological data, in contrast to the behavioral setup, where they constitute a very important variable to investigate interindividual coordination. As long as speech is concerned, the literature confirms that as only the lower jaw and vocal tract are concerned with vocalization, simple speech is compatible with fMRI acquisition provided that the rest of the head is held firmly to avoid movement. Inflatable cushions positioned on both sides of the head between the parietotemporal part of the skull and the MRI antenna (where the head rests) provide firm yet comfortable stabilization of the head.

The second difficulty is intractable and will be considered as a limitation of the study: speaking while lying supine and standing still in a noisy MRI scanner cant be considered as a natural way of speaking, even less as a natural social interaction. The use of live conversations with a human or an artificial agent will anyway provide a sufficient contrast in terms of social interaction to find differences hypothesized from the findings from more classical paradigms using simple perception or fake interactions. Surface recordings (EEG and MEG), while keeping freedom of movements and being more natural, don't allow recording the activity of deep brain structures or some cortical areas (e.g. depth of sulci), and spatial resolution is lower. In addition, muscles movement involved in speech present different challenges as they cause artifacts in these recordings.

3. Data sets, physiological pilot

The objective of the research is to characterize behavioural events that are temporally associated with physiological events. Preprocessing includes the precise synchronization of the behavioural and physiological time-series acquired independently, and the extraction of meaningful events from the time-series.

3.1. Electrodermal Activity

The example of the analysis of the electrodermal activity acquired in the behavioural pilot is used to illustrate the proposed approach more generally for the analysis of all physiological data. Using a Matlab toolbox for the analysis of electrodermal activity data (Ledalab (Benedek and Kaernbach, 2010)), the raw electrodermal recording is decomposed into phasic components and tonic responses. Tonic responses are deconvoluted in order to identify the timing and the intensity of the events responsible for each of the responses identified within the one-minute recording of each condition. The timing of the tonic electrodermal responses events is used to reconstruct a 30Hz time-series with delta functions indicating events onset (time-series [isElectrodermalEvent]).

3.2. Gaze Tracking

Eye tracking was recorded using standard procedures with from FaceLab5 from Seeing Machines. This system does not require physical constraint so that the participants remained free of their movements. Screen x and y voxel coordinates of the direction of the gaze and of the face on the screen were extracted. Eye closure and saccades were also extracted for filtering out unusable data. Time-series were downsampled from 60 to 30 Hz by decimation (removing one every other time point) to match the rest of the recorded times series. Moreover, video data was analyzed to extract facial features for each frame. A face recognition algorithm (Facial Feature Detection & Tracking; (Xiong and De la Torre, 2013)) was run frame by frame to identify the face present in the image. Screen x and y pixel coordinates of 49 keypoints on the face were extracted as well as the position

and rotation of the face mask in relation to the screen normal vector. Face tracking results were combined with gaze tracking data to provide binary 30 Hz time series indicating gaze information for each frame. First, using face tracking coordinates, the position of the face, the eyes and the mouth on the screen were calculated for each frame and used to define regions of interest. Then, using gaze tracking coordinates, 30 Hz binary time series were created indicating whether or not the gaze was within these regions of interest (is the gaze on the screen [isData], is the gaze on the face [isFace], is the gaze on the eyes [isEyes], is the gaze on the mouth [isMouth]).

3.3. Linguistic features extraction

The audio files were manually transcribed and forced-aligned at the token level with SPPAS (Bigi and Hirst, 2012). This alignment allowed us to produce time series based on IPUs (stretches of speech separated by pauses of a given duration threshold, here we used 100, 200, 400 and 800ms) and tokens. These identified IPUs were used to construct three time series describing which agent is speaking [isParticipantSpeak], [isDiscussantSpeak], [isSilent].
Transcription had been realized in standard orthographic conventions without omitting truncated words, back-channels and other spontaneous speech phenomena such as disfluencies. We also determined whether the IPU is a feedback behavior and whether it hosts disfluencies, based on the transcription content.

4. Analysis

As explained above, multiple 30Hz time series were produced during preprocessing. Physiological events are considered as temporally associated with behavioural events, from which psychophysiological co-occurences will be investigated (Bach and Friston, 2013). A probabilistic approach was chosen under the assumption that it is adapted to the ecological type of relationships expected here, which are multidimensional (speech, face and eye movements, physiology) as well as noisy given the ecological design. The probability of a given behavior, the probability of having an electrodermal response, and the probability of having both the behavior and the electrodermal response was calculated in time windows of 100, 167, 200, 333 and 500ms. The posterior probability of certain behaviours (e.g. looking at the mouth of the interlocutor) giving rise to an electrodermal response was performed with a direct application of Bayes theorem. It is particularly well suited here to take into account that events were not controlled in terms of their probability and temporal distribution given the unconstrained nature of the conversational interaction.
Here, we present the analysis of linguistic and gaze behaviors in relation with electrodermal responses. Given the deconvolution of electrodermal activity and the physiological delays, synchronicity at the frequency used for data preprocessing (30 Hz, meaning co-occurrence of events within 33ms windows) is unrealistic. An exploratory approach was adopted, choosing time windows of 100, 167, 200, 333 and 500ms to calculate co-occurences. The effect of the two experimental factors on the posterior probabilities were assessed with linear statistics (ANOVA). Figure 2 presents the effect of these factors on the posterior probabilities associated with different behaviours. For example, panel 1 indicates how the posterior probability of observing a electrodermal response when the participant listens to the other agent is affected by the Agent, the Interaction, and the interaction between Nature and Interaction. While the results were quite consistent across the sizes of time windows, 200 ms always provided, when significant, the most significant effect. It is interesting to compare this to the conclusion of (Laming, 1968) that states that a simple reaction time to a visual stimulus, when no other task is required, is around 220ms. Significant effects of the agent are presented in figure 3: both when the eyes and the mouth are being watched, the posterior probability of co-occurrence of the behavioral and physiological event is higher for the Human compared to the artificial agent. In other words, natural behaviours in a conversation, such as looking at the eyes or the mouth of the discussant, is more likely when the other agent is human compared to artificial.
Finally, we investigated the differences in terms of interactivity across the different conditions from a linguistic perspective. The quantification is currently based on the ratio of IPUs directly involving *feedback* compared to the total number of IPUs in the interaction. The different conditions are leading to significantly different ratios in the expected directions, i.e. the more natural the interaction, the more feedback related IPUs are produced proportionally. More precisely, the nature of the agent brings significant differences for both live ($p=0.004$) and video ($p=0.02$) conditions, while the nature of the interaction shows significant differences for the virtual agent ($p=0.04$) and for human agent ($p=0.04$).

4.1. Neurophysiological-Linguistic data sets

The fMRI corpus will be investigated using classical approaches relying on the General Linear Model and implemented in SPM toolbox (Statistical Parametric Mapping) and region-of-interest (ROI) approach. Comparing brain responses to human and to ECA during interaction will already offer an interesting validation of how the different dimensions of social cognition are affected by the nature of the interaction partner.
The core of the project is to use fMRI corpus to estimate the timing of cognitive events through a reverse inference from brain activity. The important processing step is to transform 4D fMRI signal into binary or delta functions time series. The methodology proposed uses a similar procedure than for the skin conductance response, namely devolution of fMRI signal using the hemodynamic response function. Raw fMRI signal presents difficulties in comparison with skin conductance that impair a direct application of the method, namely its size (number of voxels), the relatively low signal to noise ratio and the low frequency signal trends. Classical fMRI preprocessing steps will therefore be applied, such as high-pass filtering and temporal and spatial smoothing using a gaussian kernel. To recreate the conditions of analysis used in the deconvolution of skin conductance, a region of interest (ROI) approach will be chosen. These regions will be chosen based on an in depth knowledge of their contribution to social cognition. ROIs

will be identified on the basis of anatomy (e.g. (Wolfe et al., 2015) for the hypothalamus mask) and on the basis of existing coordinates and on the basis of localizer scans (eg voice localizer (Latinus and Belin, 2011)). We will obtain time-series (similar to the skin conductance time-series) for all ROIs. The result of the deconvolution will be binary (for sustained activity: is present for a certain duration) or delta function (for event-related: duration is null) time-series, identifying when during the course of the conversation do specific the cognitive events associated with the ROI analyzed (for example mentalizing for medial prefrontal cortex activity) occur. Physiological data will mainly be used as latent variables in a Dynamic Bayesian Network for identifying the timing of unidimensional (skin conductance) or multidimensional (fMRI) cognitive events during natural conversation. The outputs are therefore time-series tagging the moments when cognitive events take place during each trial of interaction with the ECA and the human.

5. Discussion

In this paper, we presented a new experimental set-up providing precisely aligned recordings of linguistic and physiological events for analysis. We have shown that the set-up allows for the recording of fine-grained linguistic phenomena, which enables an assessment of the level of interactivity of the dialogue. More crucially we have shown that the physiological measures obtained were correlated with various communicative behaviors. Therefore, we are now in the position to conduct more in-depth experiments and analyses in the field of Social Signal Processing in order to reveal temporal, and eventually causal, relationships between multi-modal linguistic behaviors and physiological activity. We also explained how we extend our pilot to work at neurophysiological level, in particular using functional magnetic resonance imaging (fMRI).

Moreover, we discuss the challenges caused by the spontaneous and interactional nature of the activity recorded. Of course the resulting analysis is somehow much more complex than for better controlled scenarios. In particular, a tricky problem that is likely to concern most of resources and projects attempting to link linguistic data with (neuro-)physiological measurements is to decide which analytical tool to use for relating the two types of data. The question of the temporal association will be also a general issue for this kind of project. The solution proposed here is only a first attempt to try to get a reliable association between the linguistic events and (neuro-)physiological signals. Finally, we illustrated the need to find the right balance set-up between what is desired, needed and required on the technical experimental side and the level of naturalness we would like to reach for our experiments. fMRI is probably among the most constraining experimental set-up, yet preliminary pilot in the scanner have shown that subjects manage to interact rather normally with someone outside the machine through our adapted communication device.

From a language sciences viewpoint, this project is a unique opportunity to correlate linguistic observations and models with other sources of evidence and in particular neurophysiological data. More precisely it allows to cross-validate verbal behaviors with neurophysiological recordings in the context of social interactions that are both spontaneous and controlled along a number of relevant dimensions thanks to the use of the artificial agent. In total, we will produce several hours of semi-controlled conversational data aligned with other behavioral information (gaze and face tracking) and physiological recordings (brain activity, skin conductance, respiration and peripheral blood pulse. We have also explored facial muscles activity and the head movements in (Ochs et al., 2017) from the participants of the behavioral experiment. This will constitute a unique corpus to further investigate the neurophysiological correlates of conversational activities allowing to address questions related to planning in interaction, face management and more generally about the interplay between cognitive, physiological and interactional constraints in language production.

6. Acknowledgments

The current work has benefited from support from the French government, managed by the French National Agency for Research (ANR), through the Brain and Language Research Institute funding (LABEX ANR-11-LABX-0036) and the convergence Institute for Language, Communication and the brain (ILCB ANR-16-CONV-0002), and the French Fund for Medical Research. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French Investissements dAvenir programme (project PhysSocial).

7. Bibliographical References

Bach, D. R. and Friston, K. J. (2013). Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, 50(1):15–22.

Benedek, M. and Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 47(4):647–658.

Bevacqua, E., Prepin, K., Niewiadomski, R., de Sevin, E., and Pelachaud, C. (2010). Greta: Towards an interactive conversational virtual companion. *Artificial Companions in Society: perspectives on the Present and Future*, pages 143–156.

Bigi, B. and Hirst, D. (2012). Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody. In *Speech Prosody*, pages 1–4.

Chaminade, T. (2017). An experimental approach to study the physiology of natural social interactions. *Interaction studies*, 18(2):254–275. Fq6kd Times Cited:0 Cited References Count:31.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.

Latinus, M. and Belin, P. (2011). Human voice perception. *Current Biology*, 21(4):R143–R145.

Ochs, M., Libermann, N., Boidin, A., and Chaminade, T. (2017). Do you speak to a human or a virtual agent? automatic analysis of user's social cues during mediated communication. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI), Glasgow, UK*, page 9 pages.

Pelachaud, C. (2015). Greta: an interactive expressive embodied conversational agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 5–5. International Foundation for Autonomous Agents and Multiagent Systems.

Roth, W. T., Dawson, M. E., and Filion, D. L. (2012). Publication recommendations for electrodermal measurements. *Psychophysiology*, 49:1017–1034.

Wolfe, F. H., Auzias, G., Deruelle, C., and Chaminade, T. (2015). Focal atrophy of the hypothalamus associated with third ventricle enlargement in autism spectrum disorder. *NeuroReport*, 26(17):1017–1022.

Xiong, X. and De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.

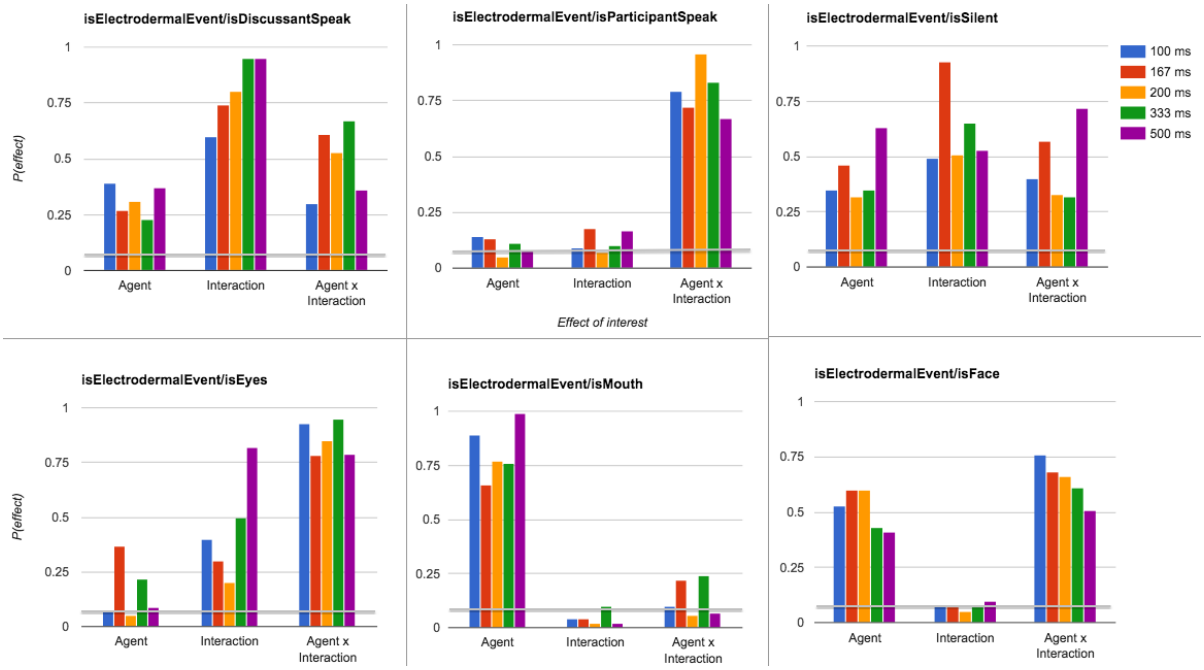

Figure 2: Outputs of ANOVAs on the effect of experimental factors Agent and Interaction on posterior probabilities of obtaining a physiological response co-occuring with a particular behaviour. In abscissa are the effects of interest (main effect "nature of the Agent (human vs artificial)", "nature of the Interaction (live vs video)", and interaction between the two factors) and in ordinate the probability (grey line: p=0.05). Colours represent the five time windows used to define IPUs.

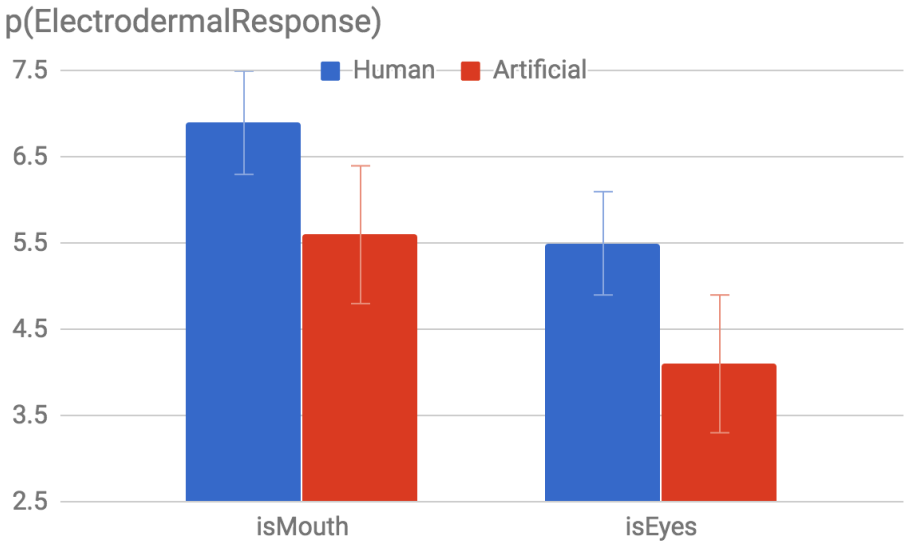

Figure 3: Posterior probability in per cent) of observing a physiological event given a behavior (participant looking at the mouth or looking at the eyes) as a function of the Agent.

A Dataset for Studying Idiom Processing with EEG

Philippe Blache, Stéphane Rauzy, Deirdre Bolger, Chotiga Pattamadilok, Sophie Dufour
Aix Marseille University, CNRS, LPL, Aix-en-Provence, France
blache@lpl-aix.fr, stephane.rauzy@lpl-aix.fr

Abstract

We propose in this paper the description of a new dataset aiming at implementing EEG experiments on sentence processing. The resource contains a set of idiomatic sentences together with the corresponding non-idiomatic control sentences. Moreover, in order to study different ERP effects for idiom processing, we also introduce in this original material controlled syntactic violations. As an application, we briefly present an EEG experiment and its results.

Keywords: Idioms, dataset, EEG, syntactic violation

1. Introduction

Studying neural correlates of sentence processing is a difficult task and requires the elaboration of specific material, in which different types of information are controlled (frequency, predictability, syntactic complexity, etc.). Many works focus on phenomena precisely associated with a position or a word, such as the analysis of semantic or syntactic violations introduced by a specific word (Kutas and Federmeier, 2011; Pulvermuller et al., 2008). It is however more complex to study larger phenomena, involving entire constructions (Fillmore, 1988; Goldberg, 2003), with effects at different positions in the sentence. This is still a scientific and methodological lock, and we need to imagine linguistic contexts in which it becomes possible to predict effects at the syntactic level instead of the lexical one. Idiomatic constructions offer such a frame: they can be identified on a word-by-word basis, but are known to be processed globally (Molinaro and Carreiras, 2010; Rommers et al., 2013; Vespignani et al., 2010; Boulenger et al., 2012).

Idioms constitute a prototypical construction (Sag et al., 2002): when recognized, the complete construction (including the meaning as well as possible restrictions on the morphology and the syntax) becomes available. In our work, we intend to analyze *brain activity* in response to a syntactic violation introduced into an idiom. We compare event-related potentials (ERP) in different conditions: idioms vs. control sentences, with or without a syntactic violation. Our goal is to test the hypothesis stipulating that the difficulty of processing the violation is compensated by the activation of the idiom.

This experiment relies on a controlled dataset, made of French sentences in which all information required for implementing such work has been controlled. This constitutes a new resource of 240 sentences, half of them containing idiomatic constructions, the other being corresponding control sentences. Different types of specific information have been encoded such as the familiarity of the idiom, its recognition point as well as information on the type of violation used for this specific study.

2. Linguistic data

A first list of 1,220 French idiomatic expressions have been created from different existing lists available on the web. From this set, a sublist of 170 idioms fulfilling different criteria (familiarity, positions of the constraints violation and its detection) have been extracted by 4 experts. For each idiom, the recognition point *RP* (the word starting from where the idiom is completely recognized) is located. Idioms with "late" *RP*, located at the end of construction, are eliminated, no place being left for introducing violation.

Idiom:	`coûter les yeux de la tête`
Word-by-word:	*to cost an arm and a leg*
Meaning:	*to be very expensive*
English equivalent:	*to cost the eyes in one's head*

Figure 1: Example

In a second stage, this list has been presented to 40 naive participants. Their task was to read the beginning of each idiom (from the first word to the recognition point) and complete them. For each idiom (I), a support sentence (SS) has been built. In order to encapsulate the idiom and to avoid specific effects at the beginning and the end of the sentence, all SS start with a proper noun in a subject position and last with a sentence complement, added after the idiom, in order to let time enough for the EEG signal we want to observe to be realized. The following example illustrates such an idiomatic construction. Starting from the idiom:

(I) `avoir une idée derrière la tête`
(*to have something in mind*)

with recognizing point:

`(RP) derrière`

we build the sentence :

(SS) `Paul a une idée derrière la tête depuis ce matin.`
(*Paul has something in mind since this morning.*)

Idiom selection: From the support sentence (SS), we created the priming stimulus which span from the beginning of sentence to the recognizing point (that is included), as illustrated in the example:

`(PSS) Paul a une idée derrière ...`

The list of the 170 priming stimulus was proposed to a cohort of students (36 for the first 120 items, 25 for the remaining 50 items). It was asked to complete the sentence (without any help). The completion results were analyzed making it possible to calculate a "*familiarity*" measure for each idiom, spanning from 0 (no students can complete correctly the priming stimulus) to 1 (the entire cohort was successful in completing the priming stimulus). The familiar-

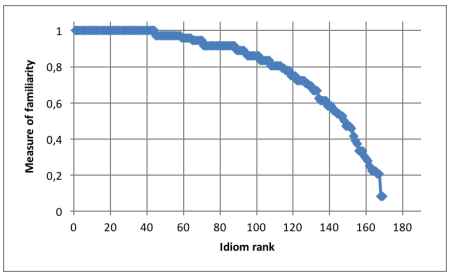

Figure 2: Measure of familiarity

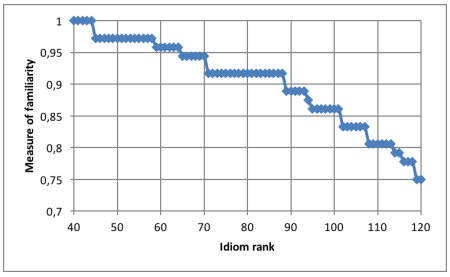

Figure 3: Familiarity of the first 120 idioms

ity measure is illustrated figure 2, the idioms being ranked by decreasing order.

In our list, 44 idioms have been successfully completed by all the participants, the first 100 idioms have a familiarity measure greater than 86% (with means familiarity of 0.96), and the first 120 idioms a familiarity greater than 75% (with a mean familiarity of 0.934) as shown figure 3.

Other idioms receive a lower measure of familiarity for different reasons. One is that the completion of the priming stimulus is ambiguous, as in the following example (which obtains a score of 0.5):

`avoir un verre dans le nez`
(*to be drunk*)

versus the non-idiomatic (but frequent) sentence:

`avoir un verre dans la main`
(*to have a glass in the hand*)

Another reason is that the idiom can be obsolete for the cohort. For example, a measure of familiarity of 0.21 is observed for the idiom:

`boire le calice jusqu'à la lie`
(*To drink from the bitter cup*)

The final list of selected idioms contains the best 120 ranked familiar idioms (with a mean familiarity measure of 0.934).

Control sentences: For each idiomatic support sentence, we created an associated *control sentence* (CS) with the same syntactic structure, the same number of words and as far possible the same lexical material. For example, to the idiomatic support sentence:

(SS) `Paul prend son courage à deux mains pour le faire`
(*Paul takes courage to do it.*)

is associated to the non idiomatic control sentence:

SS	`Paul trouve que ça lui coûte les yeux de la tête maintenant`
SS violation	`Paul trouve que ça lui coûte les yeux sur la tête maintenant`
RP	`yeux`
CS	`Paul trouve que ça lui rappelle les plats de la cantine évidemment`
CS violation	`Paul trouve que ça lui rappelle les plats sur la cantine évidemment`

Figure 4: Idiom, recognition point, control sentence

(CS) `Paul prend son paquet à deux bras pour le porter`
(*Paul takes his bundle with two arms to carry it*)

The table 4 recaps the complete set of data built from one idiom, the corresponding control sentence and the violations.

Violations: The violations are introduced either right after the recognizing point (RP+1) or two words after (RP+2). We introduced two types of violation. The first is a gender and/or number violation agreement between the determiner and the noun in the noun phrase or prepositional phrase following the recognizing point. For example, the support sentence:

(SS) `Paul a un cheveu sur la langue depuis toujours`
(*Paul has a lisp since forever*)

is associated the *violated support sentence* (VSS), in which the agreement between the determiner and the noun is violated:

(VSS) `Paul a un cheveu sur` **le** `langue depuis toujours`

Among the 120 items, there is 47 violations of this kind. The second type of violation introduced is the substitution of the preposition following the recognizing point by another one not allowed in practice, as for example:

(SS) `Paul range au fur et à mesure ses affaires`
(*Paul arranges his stuff as and when*)

(VSS) `Paul range au fur et` **en** `mesure ses affaires`

The list contains 64 items with such violation. The remaining 9 items have slightly different violation rules due to their specific syntactic structure, such as:

(SS) `Paul dit à qui veut l'entendre que c'est vrai`
Paul says to whoever wants to hear it it is true.

(VSS) `Paul dit à qui veut` **s'**`entendre que c'est vrai`

In this case, the violation concerns the accusative pronoun which was substituted by a pronominal pronoun. The same types of violation are also introduced in the control sentences.

3. EEG data

As explained above, an idiom is recognized at the *recognition point* that occurs usually 2 or 3 words after the beginning of the idiom. For example, the recognition point for the idiom "*to put all eggs in one basket*" is the noun "*eggs*". At *RP*, the entire construction is activated, making available predictions about the rest of the input. As illustrated in figure 5, scanning a new word of the input is a simple

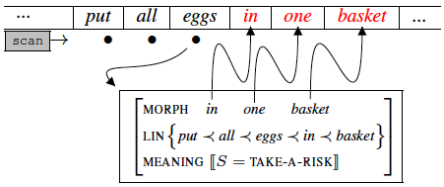

Figure 5: Processing the idiom

mechanism, matching the scanned form with the predicted one. This process remains very shallow, with no precise and in-depth unification mechanism, these words after RP being highly predicted.

One first question to be investigated is to examine whether idiomatic constructions elicit specific brain activities, and more precisely what happens before and after the *RP*. Moreover, in the violation condition, our hypothesis is that there exist compensating effects due to the construction: it is expected that the error in the idiom is identified, but not repaired.

We carried out an electrophysiological (EEG) experiment in which participants were presented with 120 French idioms (ID), 60 with violations (IDV) and 60 without (IDNV), and 120 control sentences (CTR), 60 with violations (CTRV) and 60 without (CTRNV). The stimuli were presented, word-by-word, on-screen during EEG recording. The distribution of idiom familiarity and violation type was controlled. As it is classically the case in EEG, the experiment consists in finding in the data specific electric potentials that can be associated to some stimuli. Several such potentials (called *event related potentials*) are known to be associated with language processing. For example, semantic violations are associated with a negative potential occurring 400ms after the stimulus (called N400), prediction comes with a positive potential 300ms after the stimulus (P300), etc.

		RP	*MM*	*MD1*		
IDNV	*Paul fait la pluie*	*et*	*le*	*beau*	*temps*	...
IDV	*Paul fait la pluie*	*et*	*la*	*beau*	*temps*	...
CRTNV	*Paul fait la peinture*	*et*	*le*	*gros*	*travail*	...
CRTV	*Paul fait la peinture*	*et*	*la*	*gros*	*travail*	...

Table 1: The 4 sentences generated for the idiom "*faire la pluie et le beau temps*" ("*to call the shots*") and the studied positions: the recognition point (*RP*), the modified word (*MM*) where the violation is introduced and the detection word (*MD1*) where the violation is detected for the CRTV condition (here, a gender agreement violation between the determiner and the adjective).

From the 120 sentences in their 4 conditions (idiom, idiom violated, control, control violated), we built two complementary lists of 240 items. Each list contains the whole set of 120 idioms (60 violated and 60 non-violated) and their associated 120 control sentences (60 violated and 60 non-violated). If the couple (violated idiom/non-violated control) belongs to the first list, its corresponding couple (non-violated idiom/violated control) belongs to the second list.

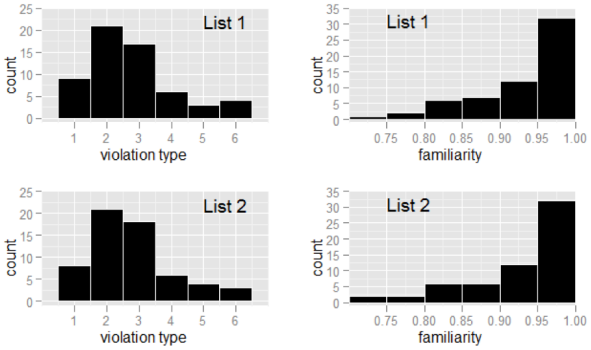

Figure 6: Repartition violation type / familiarity

The two lists have the same distribution of violation (6 different types) for each four conditions (idiom non-violated IDNV, idiom violated IDV, control non-violated CTRNV and control violated CTRV). The two lists have also the same distribution of idiom familiarity (see figure 6).

Figure 1: the distribution of the violation type (left) and of idiom familiarity (right) for the 60 non-violated idioms of the two lists. The mean familiarity is respectively of 0.9345 and 0.9338 for list 1 and list 2.

Subject material input file: The participants were split into two equal subsets, to whom one of the lists 1 or 2 is presented. One single participant can be exposed either to the non-violated idiomatic sentence and its corresponding violated control sentence or to the violated idiom and the non-violated control condition. It never happens that the same participant is asked to read the violated non-violated idiom nor violated and non-violated control sentence. For each participant, the 240 items of the list are randomized and split into six runs of 40 items. For each run, the attention of the participant is checked by inserting 4 sentence questions appealing a yes/no answer to an image presentation. The instruction is to answer *yes* if the sentence has been presented during the last run and *no* otherwise. The answers are balanced in such a way that a given subject has 12 positive and 12 negative answers to give over the experiment.

4. An EEG/ERP study on syntactic violations in idiom comprehension

Subject-level, trial-averaged EEG data was extracted for the three word positions: the *Recognition Point* (RP), the *Modified Word* (MM) where the violation is introduced and the *Detection Word* (MD1) where the violation is detected (for the CRTV condition). A two-tailed cluster-based permutation was carried out on the data for both CTRLs and IDs to compare non-violation conditions (CTRNV and IDNV) and violation conditions (CTRV and IDV).

Recognition Point (RP): As no effect of violation was expected, the violation conditions were collapsed for both CTRL and ID ((CTRNV+CTRV) vs. (IDNV+IDV)).

Statistical analyses revealed a significant ($p \leq 0.025$, two-tailed) N400 difference over centro-parietal electrodes from $\sim$390 to 550ms; CTRL presented a higher N400 ampli-

tude than ID (figure 1). N400 amplitude is generally thought to increase as a function of the difficulty of word retrieval and integration (Kutas and Federmeier, 2011). This observation is in line with previous findings of a reduced N400 in the context of idioms compared to literal sentences (Rommers et al., 2013) and is indicative of higher word-predictability at the RP for idioms compared to controls. A greater P300 effect, posited as an index of prediction processes in idioms (Molinaro and Carreiras, 2010), was observed for ID compared to CTRL. However, this did not reach statistical significance according to the cluster-based permutation test (figure 7).

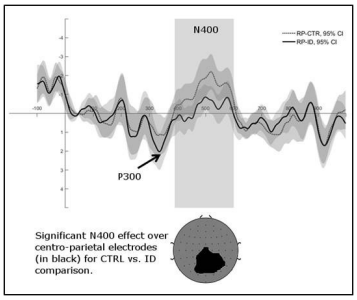

Figure 7: RP Position

Modified Word (MM): Violation effect in CTRL and ID were analyzed separately. As expected, no significant difference was revealed for CTRLs. However, for ID, IDV presented a significantly higher ($p \leq 0.025$) N400 than IDNV (figure 8).

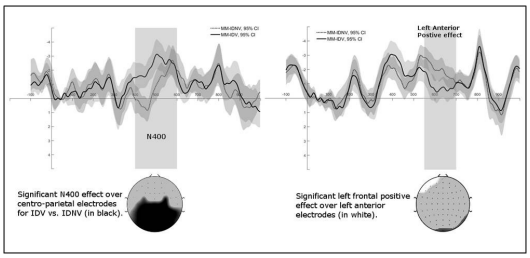

Figure 8: Position MM

This N400 effect indexes the violation of the prediction made at the RP and, so far as it indicates that the violation has already been processed, this effect also implies that, for ID, the reader is already predicting the error that will occur at MD1.

A significant difference ($p \leq 0.025$) between IDV and IDNV in the 550 to 700ms time window over left frontal electrodes (figure 2) was also revealed; IDV presented more positive-going activity compared to IDNV. This observation could be interpreted in light of (Hagoort et al., 1999) suggestion that more frontally distributed P600-like effects may reflect the over-writing of an "*active structural representation*".

Detection Word (MD1): At this position the reader detects the violation introduced at position MM for CTRL. A CTRV vs. CTRNV comparison revealed a significant N400 effect ($p \leq 0.05$) (figure 3, left); this reflects the processing of the control violation.

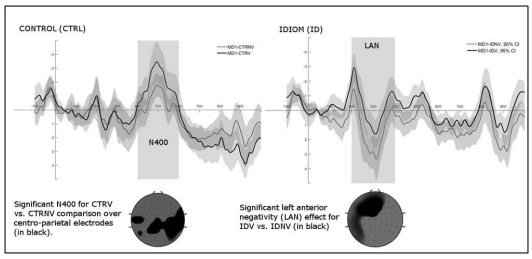

Fig. 3: MD1 position

The N400 effect is followed by increased positive-going activity for CTRV from around 600ms; this P600 effect indicates the processing of the syntactic violation. The ID condition presents a different pattern of results. The N400 was very much reduced for both IDNV and IDV and no significant N400 difference was observed as a function of the violation. However, an IDNV vs. IDV comparison revealed a significant ($p \leq 0.025$) difference over left frontal electrodes from around 200ms to 400ms, (figure 3, right) with IDV presenting more negative-going activity than IDNV. The temporal and spatial focus of this effect suggests a LAN (Left Anterior Negativity) which has been posited as reflecting syntactic processing (Friederici et al., 1996; Kluender and Kutas, 1993) rather than semantic integration.
These results validate the different hypothesis mentionned above. By showing a higher positivity (reduced N400, higher P300) starting from the recognition point (*RP*), the ERPs validate the facilitator effect after *RP* due to the prediction of the entire construction. At the modified word position (*MM*), as predicted by the model, the violation in the idiomatic construction is detected (small N400). Moreover, the unexpected element is recovered (P600), corresponding to our constraint relaxation hypothesis. The analysis of the detection word position (*MD1*) reveals clearly a specificity of violation in idiomatic constructions. In the IDV condition, the violation is already predicted starting from the modified word (*MM*). This explains the fact that no N400 occurs at this point in IDV. In the control condition, the violation is only detected at this point, which explains a high N400, followed by a repair. Finally, as predicted by the model, the violation is not repaired in IDV: the LAN at this position reveals an automatic detection of the violation, but is not followed by a repair that should have generated a P600.

5. Conclusion

The dataset presented in this paper constitutes a complete resource for the study of idiom processing. The EEG experiment done with this resource shows the compensation effect played by the idiomatic construction when faced with a syntactic violation. Such dataset opens new experimental possibilities. On top of providing a controlled material for testing hypothesis on idiom processing, it also opens directions towards new experiments in neurolinguistics for the analysis of syntactic phenomena at the sentence level. In particular, the fact that entire constructions such as idiom can be manipulated makes it possible to implement different experiments involving larger contexts than isolated words or adjacent chunks. The EEG experiment we presented is an illustration of such type of works.

6. Acknowledgements

This work has been supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

7. Bibliographical References

Boulenger, V., Shtyrov, Y., and Pulvermüller, F. (2012). When do you grasp the idea? meg evidence for instantaneous idiom understanding. *NeuroImage*, 59(4):1–12.

Fillmore, C. J. (1988). The mechanisms of "construction grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Friederici, A., Hahne, A., and Mecklinger, A. (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1219–1248.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Hagoort, P., Brown, C. M., and Osterhout, L. (1999). The neurocognition of syntactic processing. In C. M. Brown et al., editors, *The neurocognition of language*. Oxford University Press.

Kluender, R. and Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, 8(4):573–633.

Kutas, M. and Federmeier, K. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *The Annual Review of Psychology*, 62(1):621–647.

Molinaro, N. and Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biological Psychology*, 83(3):176–190.

Pulvermuller, F., Shtyrov, Y., Hasting, A. S., and Carlyon, R. P. (2008). Syntax as a reflex: Neurophysiological evidence for early automaticity of grammatical processing. *Brain and Language*, 104(3):244–253.

Rommers, J., Dijkstra, T., and Bastiaansen, M. (2013). Context-dependent Semantic Processing in the Human Brain: Evidence from Idiom Comprehension. *Journal of Cognitive Neuroscience*, 25(5):762–776, May.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., and Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of Cognitive Neuroscience*, 22(8):1682–1700.

Ontology and Synesthesia: Language, Sense and the Conceptual Inventory

Adam Pease, Chu-Ren Huang
Infosys, The Hong Kong Polytechnic University
Palo Alto, Hong Kong
adam.pease@infosys.com, churen.huang@polyu.edu.hk

Abstract

We examine the ontological evidence for synesthesia. We employ the Suggested Upper Merged Ontology (SUMO), which has a complete set of manual mappings from its terms to the lexical elements in Princeton WordNet. By looking at polysemous words that map to SUMO terms that address more than one human sensory modality, we attempt to provide an inventory of concepts. We compare this list to prior work in creating corpora of such words and concepts built exclusively for the purpose of this sort of study.

1. Extended Abstract

Human language provides some evidence for linking among the human senses. One can talk about a *sharp* object and a *sharp* taste. Light can be *bright* and so can sound. We attempt to provide an inventory of such usage. Previous work attempting to do so (Strik-Lievers and Huang, 2016) (Strik-Lievers and Winter, 2017) has relied on using human annotators to assess word lists. Using a previously built, very large ontology, makes quicker and possibly more comprehensive work of collecting relevant lexical items.

Note that we are not concerned with all types of synesthesia here, such as a link between numbers and colors (Van-Bergeijk, 2010), but only between adjectives that can be applied to more than one human sense (as well as thoughts and emotions). We might expect that the most common forms of synesthesia would be likely to have the largest number of words applying to both senses, and this appears sometimes to be borne out in our study. Chromasthesia (Cytowic and Eagleman, 2009) is relatively common and there are a significant number of words that describe both sight and sound. Interestingly however, although sound-touch synesthesia is rare (Naumer and van den Bosch, 2009), the richest set of words is found in this category.

For this work, we rely on the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001; Pease, 2011)[1], which is a hand-built, open source ontology defined in higher order logic (Benzmüller, 2015). SUMO has also been linked, in an entirely manual process, to all of the approximately 117,000 word senses in the WordNet (Fellbaum, 1998) lexicon (Niles and Pease, 2003). Unlike taxonomies or semantic networks, the semantics of SUMO are defined logically, rather than with recourse to human understanding of the labels of nodes - the semantics are in the formal programming language constructs and the semantics of the program do not change even if all the labels are changed. As such, SUMO is suitable as an *interlingua* (Pease and Fellbaum, 2010) and is linked via Open Multi-Lingual Wordnet (Bond et al., 2014) to several dozen human languages. SUMO has been created over a 17-year period and has roughly 20,000 terms and 80,000 logical statements about those terms.

SUMO has an extensive hierarchy of processes, including those that relate to the five human senses. These are the "seed" concepts that we use to begin our exploration. They are **Tasting**, **Hearing**, **Smelling**, **Seeing** and **Touching**. These in turn are related to a hierarchy of **Attribute**s appropriate to their respective senses, as in the rule

```
(=>
  (and
    (instance ?TASTE Tasting)
    (patient ?TASTE ?OBJ))
  (exists (?ATTR)
    (and
      (instance ?ATTR TasteAttribute)
      (attribute ?OBJ ?ATTR))))
```

This says, in first order logic, that if there is a **Tasting** process then the **patient** (or object) of the tasting has a particular **TasteAttribute**. A portion of the **Attribute** hierarchy is as follows

```
RelationalAttribute
    PerceptualAttribute
        SoundAttribute
            Stressed
            Audible
            Inaudible
        TasteAttribute
            Sweetness
            Bitterness
            Sourness
            Saltiness
            UmamiTaste
        OlfactoryAttribute
        VisualAttribute
            ColorAttribute
                SpectralColor
                SecondaryColor
                PrimaryColor
        TextureAttribute
            Smooth
            Rough
```

To the list of **PerceptualAttribute**s that are relevant for synesthesa, we also add the SUMO concepts **RadiatingSound**, **Music** and **RadiatingLight**.

2. Lexical Inventory

SUMO does not contain an exhaustive list of perceptual concepts, and it may not even be possible to create such

[1] http://www.ontologyportal.org

a complete list. But it can be improved and extended, and that can be a byproduct of the current exploration. There are several other attributes that are not specific to any one sense. In fact, these may be an appropriate initial focus for investigation. They are **TemperatureAttribute** and **ShapeAttribute**.

Let's look at the word *warm*[2]. We find 10 adjective senses in WordNet that have mappings to different SUMO terms. Of some interest is that many of the senses relate to emotional states as well as sensory information. We should note that adjectives as isolated lexical elements often have very little semantics but rather derive their meaning from their relationship to other lexical elements. As such, several senses have the unsatisfying mapping to a SUMO **SubjectiveAssessmentAttribute**. Note also that language is less precise than mathematical logic, so most mappings are approximate and state that a more specific notion in WordNet is mapped to a more general notion in SUMO

- characterized by strong enthusiasm; "ardent revolutionaries"; "warm support": **EmotionalState**
- having or displaying warmth or affection; "affectionate children"; "a fond embrace"; "fond of his nephew"; "a tender glance"; "a warm embrace": **SubjectiveStrongPositiveAttribute**
- easily aroused or excited; "a quick temper"; "a warm temper": **PsychologicalAttribute**
- of a seeker; near to the object sought; "you're getting warm"; "hot on the trail": **Near**
- characterized by liveliness or excitement or disagreement; "a warm debate": **SubjectiveAssessmentAttribute**
- uncomfortable because of possible danger or trouble; "made things warm for the bookies": **SubjectiveAssessmentAttribute**
- psychologically warm; friendly and responsive; "a warm greeting"; "a warm personality"; "warm support": **SubjectiveWeakPositiveAttribute**
- (color) inducing the impression of warmth; used especially of reds and oranges and yellows; "warm reds and yellows and orange": **ColorAttribute**
- having or producing a comfortable and agreeable degree of heat or imparting or maintaining heat; "a warm body"; "a warm room"; "a warm climate"; "a warm coat": **WarmTemperature**
- freshly made or left; "a warm trail"; "the scent is warm": **SubjectiveWeakPositiveAttribute**

From this list we see that *warm* shows evidence of synesthesia in that it maps both to **ColorAttribute** (which is a subclass of **VisualAttribute**) as well as **WarmTemperature** (which is a subclass of **TemperatureAttribute**).

We should emphasize that SUMO, as a comprehensive ontology, contains information about concepts that is much broader than any one study such as this. It provides a common framework for linking the diverse information necessary as a basis for computers to understand and reason with information about the world. For example, **WarmTemperature** is not only linked to the word *warm*, but defined as a **TemperatureAttribute**, that is related to and the successor to other adjectival concepts like **CoolTemperature** via the relation **successorAttribute** and to other common sense notions like that a functioning heated swimming pool will have the attribute of being warm.

```
(=>
  (and
    (instance ?X HeatedPool)
    (contains ?X ?WATER)
    (instance ?WATER Water)
    (part ?X ?HEATER)
    (instance ?HEATER WaterHeater)
    (attribute ?HEATER DeviceOn))
  (attribute ?WATER WarmTemperature))
```

3. Inventory Differences

We compared Strik Lievers & Huang's list of 406 words (Strik-Lievers and Huang, 2016) and those in SUMO and WordNet (hereafter "SL&H"). The SUMO-WordNet corpus finds many more candidate synesthesia words. This is to be expected since we are taking inventory of terms in a dictionary, rather than asking a small group of people to come up with words just for the purpose of a study. There were some words however found in SL&H not found in SUMO's initial list of synesthesia words. These are potentially more interesting.

Let's look at a few examples:

- *translucent* has only one sense in WordNet, which explains its absence from SUMO's list of synesthesia words. Dictionary.com shows additional senses pertaining to clarity of thought, although not related to the other human senses. The first 50 examples returned by the Corpus of Contemporary American English[3] also appear to show just the visual sense.
- *gloat* has WordNet senses pertaining to an emotional state as well as a kind of **Communication** but not different perceptual senses.
- *banjo* has only one sense in WordNet and Dictionary.com, pertaining to the instrument.
- *sunny* has both the literal meaning of light from the sun as well as the emotional disposition.

While just a sample, these examples appear to indicate that SL&H may include some words that have limited evidence of synesthetic usage although if one allows metaphorial senses of emotion or thought then

[2] http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?word=warm&POS=0

[3] https://corpus.byu.edu/coca/

there is more evidence. To examine this theory we added **PsychologicalProcess** and **EmotionalState** to check for overlap of these terms (and their associated word senses), and terms for the five human senses, and this did considerbly expand our results.

This examination has also highlighted some SUMO-WordNet mappings that needed improvement. For example, the WordNet entry for *translucent*:

- allowing light to pass through diffusely; "translucent amber"; "semitransparent curtains at the windows". SUMO Mappings: **SubjectiveAssessmentAttribute** (subsuming mapping)

We corrected the mapping to be **VisualAttribute**, which then solved the problem of *translucent* appearing in SL&H but not in SUMO's list of candidate synethesia terms, although to qualify as a synesthesia word we would also need the sense of "clarity of thought" to be added to WordNet. A simpler case of a SUMO-WordNet error is *sour* where there is a link to another subjective attribute which could be made more specific by linking to an **EmotionalState**. In fact, since the emotion ontology in SUMO is relatively new, there are a number of such words that haven't been linked to the new emotion ontology terms. In general, adjectives and adverbs have received less attention in the ongoing SUMO-WordNet linking project than the nouns and verbs.

We added the concept of **MusicalInstrument** to our list to cover *banjo* and other instruments. There were also a few more obscure instruments found in SL&H and not in WordNet (*castanet*, *cithara* and *pianola*), so we added them to our lexicon by defining the terms and their lexical entries in SUMO. Note that these are simply candidate terms that have some sensory association, but are not necessarily synesthetic words. In the end, we wind up with a small set of SUMO concepts to cover the five human senses plus thought and emotion. We also need a general category for "perception" concepts that are not further classified. The full set is in Figure 1.

Compared to SL&H we find the following metrics (see Figure 2) for a list built from SUMO and WordNet. We compare the initial analysis with just the five human senses with an expanded list that adds terms pertaining to thought and emotion ("with t&e") as well as perception generally. The first row shows all of the words that have evidence of synesthesia in SUMO - each word appears associated with multiple human sense concepts (as well as those for thoughts and emotions in the second column). The next row shows the full inventory of SUMO sensory concepts, and is comparable to the list of 406 words developed by hand in SL&H, but of course much larger. In the following rows we show the intersection and difference between the SUMO-derived list of words and that of SL&H. "overlap with SL&H" is the set of words found in SL&H. After iterating on correcting some SUMO-WordNet mappings, and expanding the set of SUMO seed concepts, we arrive at only 71 words from SL&H that are not found. They show no evidence of lexically- or ontologically-justified synesthesia but may be the result of rare metaphorical uses in corpora. Further investigation is needed.

taste	**Tasting** **TasteAttribute**
sound	**Hearing** **RadiatingSound** **Music** **MusicalInstrument** **MusicGenre** **MusicalGroup** **SoundAttribute**
smell	**Smelling** **OlfactoryAttribute**
sight	**Seeing** **RadiatingLight** **VisualAttribute**
touch	**Touching** **TextureAttribute** **TemperatureAttribute** **ShapeAttribute**
perception	**PerceptualAttribute**
thought	**PsychologicalProcess** **PsychologicalAttribute**
emotion	**EmotionalState**

Figure 1: SUMO Terms

	initial	**with t&e**
synesthesia words	149	3017
SUMO candidates	5405	11155
overlap with SL&H	320	335
SUMO not in SL&H	5085	10825
SL&H not in SUMO syn.	367	149
SL&H not in SUMO	86	71

Figure 2: Word Statistics

A sample of the list of concepts that appear to pertain to more than one human sense are as follows, with each bracketed list of words prefixed by the two senses to which they pertain. The full list is on line at the URL listed in the Appendix below.

emotion : taste [*keenness, hotness, hot*]
emotion : sound [*cheer, bright, low, strain, tumult, high, ...*]
emotion : sight [*ardent, black, warm, shadow, bright, livid, beaming, ...*]
emotion : touch [*mushy, keenness, wound, jar, kick, boot, itch, ...*]
thought : taste [*keenness, savour, dry, nutty, blandness, ...*]
thought : sound [*click, hang, laugh, hark, motive, ...*]
thought : touch [*pick, hang, connect, keenness, projection, ...*]
thought : smell [*whiff, smelling, odour, snuff, scent_out, get_a_whiff, ...*]
taste : sound [*acrid, pungent, flat, sweet, sour, bitter*]
taste : sight [*sharpness, hot, flat, gingery, rich*]
taste : touch [*nip, sharpness, crisp, flat, smack, acuteness, coarseness, nutlike, keenness, bite*]
taste : smell [*sweetness, sweet, sour, acridity*]

sound : sight [*projection, pink, peep, reverberate, colouration, undertone, colour, light, bright, silvern, ...*]
sound : touch [*scratch, thud, wind, hang, roll, pat, projection, retroflex, tweet, ping, lap, pipe, ...*]
sound : smell [*wind, sour, sweet, whiff, high*]
sight : touch [*halo, projection, catching, flick, dull, flare, radial, pearl, radiate, shot, ...*]
sight : smell [*snuff*]
touch : smell [*wind, nose*]

4. Conclusions and Future Work

Using SUMO and WordNet can provide a more efficient way to collect terms that provide linguistic evidence of synesthesia.

Future work should also be able to examine the correspondence of these senses in English and the lexical inventory of Open Multi-lingual Wordnet, since SUMO is linked to OMW as well as Princeton's English WordNet.

Another possible experiment would be to take the list of sensory words from the SUMO analysis and look for corpus data that shows synesthesia, similar to (Strik-Lievers and Huang, 2016) or by looking for types or concepts, rather than simply words, that participate, as in (Pease and Cheung, 2018).

We also should be able to link and align this resource with neuro-cognitive experimental information. The ontology can play a role in providing a framework of linking of heterogeneous data. We hope also linking to behavioral data such as modal exclusivity data (Lynott and Connell, 2009) (Chen et al., 2017).

5. Bibliographical References

Benzmüller, C. (2015). Higher-order automated theorem provers. In David Delahaye et al., editors, *All about Proofs, Proof for All*, Mathematical Logic and Foundations, pages 171–214. College Publications, London, UK.

Bond, F., Fellbaum, C., Hsieh, S., Huang, C., Pease, A., and Vossen, P. (2014). A multilingual lexico-semantic database and ontology. In *Towards the Multilingual Semantic Web, Principles, Methods and Applications*, pages 243–258.

Chen, I.-H., Zhao, Q., Wang, S., Long, Y., and Huang, C.-R. (2017). Exclusivity and competition of sensory modalities: Evidence from mandarin synaesthesia. In *International Cognitive Linguistics Conference, At Tartu, Estonia*.

Cytowic, R. and Eagleman, D. (2009). *Wednesday is Indigo Blue: Discovering the Brain of Synesthesia*. Bradford books. MIT Press.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Lynott, D. and Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41(2):558–564, May.

Naumer, M. J. and van den Bosch, J. J. F. (2009). Touching sounds: Thalamocortical plasticity and the neural basis of multisensory integration. *Journal of Neurophysiology*, 102(1):7–8.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In Chris Welty et al., editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416.

Pease, A. and Cheung, A. K.-F. (2018). Toward a semantic concordancer. In Francis Bond, et al., editors, *Proceedings of The 9th Global WordNet Conference*.

Pease, A. and Fellbaum, C. (2010). Formal Ontology as Interlingua: The SUMO and WordNet Linking Project and GlobalWordNet. In Chu-Ren Huang, editor, *Ontologies and Lexical Resources*. Cambridge University Press, Cambridge.

Pease, A. (2011). *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Strik-Lievers, F. and Huang, C.-R. (2016). A lexicon of perception for the identification of synaesthetic metaphors in corpora. In *the Language Resources and Evaluation Conference (LREC-2016)*.

Strik-Lievers, F. and Winter, B. (2017). Sensory language across lexical categories. *Lingua*.

VanBergeijk, E. (2010). Daniel tammet: Born on a blue day: Inside the extraordinary mind of an autistic savant. *Journal of Autism and Developmental Disorders*, 40(10):1293–1293, Oct.

A Supplemental Material

The software and ontology are available under GNU GPL license at `https://github.com/ontologyportal`. The code for computing the list of synesthetic terms and the differences between SUMO's list of sensory terms and those in Strik Liever's compilation is found in the Java method **`com.articulate.sigma.WordNetUtilities.synesthesiaCompare()`** . The full list of synesthesia words and statistics are on line at `http://www.ontologyportal.org/synesthesia.txt`

Can Eye Movement Data Be Used As Ground Truth For Word Embeddings Evaluation?

Amir Bakarov
The National Research University Higher School of Economics, Moscow, Russia
Federal Research Center 'Computer Science and Control' of Russian Academy of Sciences, Moscow, Russia
amirbakarov@gmail.com

Abstract

In recent years a certain success in the task of modeling lexical semantics was obtained with distributional semantic models. Nevertheless, the scientific community is still unaware what is the most reliable evaluation method for these models. Some researchers argue that the only possible gold standard could be obtained from neuro-cognitive resources that store information about human cognition. One of such resources is eye movement data on silent reading. The goal of this work is to test the hypothesis of whether such data could be used to evaluate distributional semantic models on different languages. We propose experiments with English and Russian eye movement datasets (Provo Corpus, GECO and Russian Sentence Corpus), word vectors (Skip-Gram models trained on national corpora and Web corpora) and word similarity datasets of Russian and English assessed by humans in order to find the existence of correlation between embeddings and eye movement data and test the hypothesis that this correlation is language independent. As a result, we found that the validity of the hypothesis being tested could be questioned.

Keywords: word embeddings, eye-tracking, distributional semantics, evaluation

1. Introduction

Dense vector word representations (*word embeddings)* are gaining popularity in the scientific community because of their proved efficiency in certain downstream task. However, the we are still unaware what is the most reliable evaluation method for these models. Construction of a dataset for certain downstream tasks is expensive, and some researchers argue that it is better to have a universal dataset on which can be evaluated the ability of task independent word vectors to model the structure and space of lexical semantics (or different aspects of semantics) (Schnabel et al., 2015).

Evaluation on such data is usually called *intrinsic evaluation* (as an opposite to the *extrinsic evaluation* which is an evaluation on downstream tasks). However, it is not actually clear how to evaluate modeling of lexical semantics. Different intrinsic evaluation tasks propose different notations of semantics, but most of the existing methods of intrinsic evaluation (like *word similarity* method or *word analogy* method) are criticized (Batchkarov et al., 2016; Rogers et al., 2017). Thus, some researchers started to propose experimental evaluation techniques trying to use *neuro-cognitive resources* as a gold standard for semantic modeling.

Among these techniques, there are methods of detecting neural activation patterns on text processing (like *magnetoencephalography* (Wehbe et al., 2014)), methods of detecting reaction time on reading (like *semantic priming* (Auguste et al., 2017)), and many others. In this work, we propose another possible data source which is *eye movement on silent reading.* This data consists of measurements of time of gaze fixation on each word, amount of returns of the gaze, and so on. Eye movement data is supposed to be related to human language processing and possibly stores information about lexical semantics, so hypothetically the correlation with word embeddings could be found. Similar experiments with English eye movement data were already proposed in (Søgaard, 2016), and the results did not show strong correlation. However, it is not clear whether the same outcome will be observed in other languages. If eye movement data is language dependent, that this could support the idea that those data are less related to the semantics. So, the *main aim of this paper* is to answer the question whether word embeddings in one language correlate with eye movement data in another language (and how much). Of course, we cannot talk about investigating language independence (for that we would need many more typologically diverse languages), but possibly we could highlight some features of this comparison investigation two distinct languages.

This study proposes experiments with eye movement data and word embeddings for Russian as 'another' language (we consider Russian because we are native speakers of this language, so we are able to interpret the obtained results), trying to answer the question of whether word embeddings in one language correlate with eye movement data in another, and whether the correlations between two language specific embeddings and eye movement data are comparable. Probably, distributional semantics and eye movement data process different types of semantics, so we could expect that the correlation would be low.

This paper is organized as follows. Section 2 puts our work in the context of previous studies. Section 3 describes the data that we are using in the current work. Section 4 is about the experimental setup. The results of the comparison are reported in Section 5, while Section 6 concludes the paper.

2. Related work

The basic idea of this work is related to the hypothesis that observable features of human text processing (like **the time of reading of a certain word**) are based not only on the surface features of a linguistic sign but also on its meaning.

Target word	The nearest neighbor word		
	RSC, *gaze*	**Araneum**, *embeddings*	**Ruscorpora**, *embeddings*
mikrob (*microbe*)	lekarstvo (*cure*)	boleznetvorniy (*pathogenic*)	mikroorganizm (*microorganism*)
speciya (*spice*)	kuriniy (*chicken*)	priprava (*seasoning*)	speciya (*spice*)
zanyatie (*class*)	student (*student*)	trenirovka (*training*)	klassniy (*in class*)
plutovat (*to cheat*)	ministr (*minister*)	tyrit' (*to steal*)	vorovat (*to steal*)
chaska (*cup*)	vedro (*bucket*)	stakan (*glass*)	chaynik (*kettle*)

Table 1: Top-1 nearest neighbors for Russian gaze vectors (first column) and Russian embeddings vectors (second and third columns). The first word is transliteration in Russian, the word in brackets is translation into English.

There are certain studies based on eye movement data that prove correlation between word semantics and reading time (Smith and Levy, 2013; Hohenstein, 2013). More precisely, while reading a human's brain continuously builds a model of context for already read words and integrates each new word in context, comparing it with contexts stored in the memory. The effort of this integration is inversely proportional to how probable the word is, so when an encountered word is highly unpredictable, then the time of its reading should increase.

This research is strongly inspired by (Søgaard, 2016), which evaluated word embeddings against eye movement data from the Dundee Corpus (Kennedy et al., 2003) using aggregate statistics of eye movement data features. Their experiments showed that there is no notable agreement between eye movement data and word embeddings. However, these experiments were performed for English only. In our work we extend the propose of (Søgaard, 2016), making similar experiments for English and Russian to test the hypothesis whether the same outcome will be observed in other languages.

3. Datasets

3.1. Eye movement data

Eye movement data is obtained through *eye-tracking*, the process of measuring the point of the human gaze on the screen. In other words, when a person reads text on the screen, a special mechanism called eye-tracker tracks the movement of the gaze and records the information about the reading. A number of different features can be recorded, e.g. how long the gaze was fixated on a certain word, how many times the gaze returned back, and so on. We averaged the data of all examinees to obtain a feature vector for each words, so in these vectors each component would report value of each of the tracked features. We normalized these values, obtaining values of each of the features ranging from -1 to 1, and used vectors of normalized eye movement features as a gold standard for word embeddings evaluation (here and later we will us the notion of **gaze vectors**).

Russian. As a source of Russian eye movement data we used *Russian Sentence Corpus (RSC)* (Laurinavichyute et al., 2017) which contains data about reading 144 Russian sentences by 96 native speakers. After averaging examinees scores for each token and averaging word form scores for each lemma, we obtained a dataset with information on eye movements for 701 single words.

English. We used an English eye movement corpus which is the *Provo Corpus* (Luke and Christianson, 2017) (the *Dundee Corpus* used in (Søgaard, 2016) is not publicly available). It contains data on reading 55 English paragraphs by 84 native speakers. We obtained vectors with information on eye movements for 1185 words from this data (with the manipulations described above). We are also aware of another publicly available English corpus, *Ghent Eye-Tracking Corpus (GECO)* (Cop et al., 2017) containing data on reading 5 000 English sentences by native and bilingual (for which English is a second language) English speakers (33 participants overall). In this paper the data of native speakers only is used. We obtained gaze vectors for 987 words.

In all cases the raw data consisted of 17 logical and continuous eye movement features. The features included:

1. *dwell time* (summation of the duration across all fixations) on the current interest area;
2. duration of the first fixation event that was within the current interest area;
3. dwell time of the first run within the current interest area;
4. number of all fixations in a trial falling in the first run of the current interest area;
5. total fixations falling in the interest area;
6. whether the first fixation in the interest area N was preceded by a fixation in the interest area $N-1$;
7. whether the current interest area received at least one regression from the later interest areas;
8. whether regression(s) was/were made from the current interest area to the earlier interest areas;
9. dwell time from when the current interest area is first fixated until the eyes enter an interest area;
10. dwell time of the second run of fixations within the current interest area;
11. no fixation occurred in the first pass reading;
12. the duration of the first fixation made on the interest area $N+1$ after leaving the interest area N in the first pass;
13. landing position in the word of the incoming saccade;
14. direction of the incoming *saccade* (fast jump from one eye position to another);
15. character that was fixated by the incoming saccade;
16. whether the word was fixated once;
17. whether the word was fixated two or more times.

In all the datasets, the recordings were obtained with an *Eyelink 1000 Plus* desktop mount eye-tracker with a chin rest and a screen on which the sentences were presented.

3.2. Word embeddings

As a distributional model, we use *Skip-Gram*, a neural predictive algorithm that updates values on the input layer (word embeddings), trying to maximize word prediction probability by minimizing the loss of the softmax function (Mikolov et al., 2013). The reason why we employed *Word2Vec* is that it is very common in natural language processing research, evaluated and explored in many papers (note though that (Søgaard, 2016) used another type of embeddings, namely `SENNA` embeddings (Collobert et al., 2011); however, the paper describing `SENNA` propose 4 different embedding architectures, and we are not aware which exactly architecture has been used).

Russian. We used a model trained on a POS-tagged *National Russian Corpus* (we will further use the term **Ruscorpora** further) with 195 071 words in the model's vocabulary (`ruscorpora_upos_skipgram_300_5_2018`) and a model trained on a POS-tagged *Araneum Russicum Maximum* (**Areneum**) with 196 620 words in the vocabulary (`araneum_upos_skipgram_300_2_2018`) (Kutuzov and Kuzmenko, 2016). The models are available on *Rusvectores* repository[1].

English. We used a model trained on a **Google News** corpus with 2 883 863 words in the vocabulary and a model trained on a POS-tagged British National Corpus (**BNC**) with 163 473 words in the vocabulary (Fares et al., 2017). The models are available on *Nordic Language Processing Laboratory* repository[2]. All used word vectors, both Russian and English, had the same vector dimensionality of 300.

3.3. Word similarity data

Word similarity task is the most ubiquitous technique for word embeddings evaluation. Given words a and b, the task is to find scalar value reporting semantic distance between them. This task is strongly criticized in NLP community, and different researchers address problems like the obscurity of the notion of semantics, subjectivity of human judgments, and so on (Batchkarov et al., 2016).

The word similarity datasets actually differ in the types of human assessments: some datasets are assessed according to semantic similarity relation (which is commonly interpreted as a synonymy, like in words *mug* and *cup*), while other datasets are assessed by semantic relatedness (which is interpreted as co-hyponymy, like in words *cup* and *coffee*).

English. We are aware of more than 7 datasets for word similarity available for English, but in order to propose a fair comparison with gaze vectors we need to drop words that are absent in eye movement data vocabulary. To this end, we did not used datasets like *Verb-143* (Baker et al., 2014), *YP-130* (Yang and Powers, 2006), *RG-65* (Rubenstein and Goodenough, 1965) and *MC-30* (Miller and Charles, 1991), because the amount of remaining word pairs (after dropping) was too low (lower than 5). We used the only 3 datasets with assessments by semantic similarity.

1. *SimVerb-3500* (234 word pairs remained for GECO, 75 word pairs remained for Provo) (Gerz et al., 2016),
2. *SimLex-999* (37 pairs remained for GECO, 41 pairs remained for Provo) (Hill et al., 2016),
3. *WordSim-353-Similarity* (WS353-Sim) (5 word pairs remained for both GECO and Provo) (Agirre et al., 2009).

We also used 3 English datasets assessed by semantic relatedness, excluding certain datasets for the reasons described above (MTurk-287 (Radinsky et al., 2011)):

1. *MEN* (22 pairs remained for GECO, 77 pairs remained for Provo) (Bruni et al., 2014),
2. *MTurk-771* (7 pairs remained for GECO, 19 pairs remained for Provo) (Halawi et al., 2012),
3. *WordSim-353-Relatedness* (WS353-Rel) (5 pairs remained for GECO, 11 pairs remained for Provo) (Agirre et al., 2009),

Russian. The amount of word similarity datasets available for Russian is lower. All datasets we are aware of are translated versions of English datasets: SimLex-999, WordSim-353, RG-65 and MC-30 (Panchenko et al., 2016). The two latter were excluded from our comparison according to the reasons described above. So, we used the following datasets:

- The revised version of *SimLex-999*, dubbed *RuSimLex-965* (21 word pairs remained) (Kutuzov and Kunilovskaya, 2017) and translated versions
- The translated versions of *WS353-Sim* (5 pairs remained) and *WS353-Rel* (7 pairs remained) (Panchenko et al., 2016)

4. Experimental Setup

To this point, we obtained two datasets with vectors (gaze vectors and word vectors) and one dataset with scalar values (human judgments on word similarity). For each word pair in each vector dataset, we computed cosine distance between the vectors corresponding to the words in a pair. Then for every word pair in each dataset we had a float in $\{0, 1\}$ reporting similarity between two words in this pair. In the end, to find correlation of the datasets we computed Spearman correlation value for distances between word pairs. The results of that comparison are presented in the following section. The code to reproduce the experiments as well links to the datasets and models are available at out GitHub[3].

5. Results and discussion

First of all, we evaluated the gaze vector space by analyzing the nearest neighbors to certain target words. Table 1 reports the closes neighbors for Russian gaze vectors and Russian word embeddings for randomly used words (closest vectors were found with a *KD-tree* neighbor search (Maneewongvatana and Mount, 1999)). Notably, according to the notion of word relatedness, some words produced by

[1] `http://rusvectores.org/en/models/`

[2] `http://vectors.nlpl.eu/repository/`

[3] `https://github.com/bakarov/subconscious-embeddings/tree/master/eye-tracking/lincr2018`

		Similarity			Relatedness		
		SimVerb	**SimLex**	**WS353-Sim**	**MEN**	**MTurk**	**WS353-Rel**
English	**Provo**, *gaze*	0.01	-0.09	-0.2	-0.09	0.19	0.07
English	**GECO**, *gaze*	0.06	-0.28	-0.8	-0.19	0.14	-0.6
English	**Google News**, *embeddings*	$\mathbf{0.36}^{\dagger}$	$0.35^{\dagger}$	$\mathbf{0.77}^{\dagger}$	$\mathbf{0.77}^{\dagger}$	$0.67^{\dagger}$	$\mathbf{0.64}^{\dagger}$
English	**BNC**, *embeddings*	$0.18^{\dagger}$	$0.25^{\dagger}$	-	$0.76^{\dagger}$	$\mathbf{0.69}^{\dagger}$	-
Russian	**RSC**, *gaze*	-	0.24	-0.1	-	-	0.05
Russian	**Araneum**, *embeddings*	-	$\mathbf{0.39}^{\dagger}$	$0.57^{\dagger}$	-	-	$0.61^{\dagger}$
Russian	**Ruscorpora**, *embeddings*	-	$0.28^{\dagger}$	$0.74^{\dagger}$	-	-	$0.57^{\dagger}$

Table 2: Performance of English word embeddings and gaze vectors across word similarity and word relatedness tasks in Spearman's correlation value. Daggers report $pval < 0.01$, absence of a symbol report $pval > 0.05$.

		English		Russian	
		BNC	**GN**	**Ara**	**RC**
English	**GECO**	$\mathbf{0.99}^{\dagger}$	0.14	-	-
English	**Provo**	0.97	$\mathbf{0.23}^{\dagger}$	-	-
Russian	**RSC**	-	-	$\mathbf{0.65}^{\dagger}$	$\mathbf{0.63}^{\dagger}$

Table 3: Correlation of English gaze vectors with English word embeddings and Russian gaze vectors with Russian word embeddings (values report Spearman's correlation).

gaze vectors seem to be very related to target words. So, we cannot say that the gaze vectors work well, but we also not conclude that they are just random.

Table 2 reports the correlation values for Russian and English gaze vectors (as well as word vectors) with word similarity assessments. In general, the experiments show the lack of correlation even between similarity judgments and gaze vectors, giving in most cases low correlation score. Apart from the issue of embedding evaluation, this raises the problem of whether semantic similarity is a factor affecting gaze variables during reading at all. However, due to the low size of remained word pairs in preprocessed datasets, the statistical significance of obtained results could be question, so we are not able propose any confident conclusions.

The results of pairwise comparison of distances between English gaze vectors for both eye movement datasets and embeddings vectors for both corpora are presented in Table 3 (all $pval < 0.01$). The results report very high correlation with a *BNC* model and low correlation with Google News despite the fact that a *Google News* model showed better results on word similarity task. On the other hand, variation in correlation scores for two English distributional models is high, while variation for Russian model is low. This fact possibly proves our hypothesis that eye movement data behaves differently for different languages. It is also interesting that the *Araneum* model reports the highest correlation with gaze vectors, and it also has the best results among Russian embeddings on most of word similarity tasks, while English model show an inverse pattern.

To this end, we do not say that eye movement data is worth being used as a gold standard for evaluation since we are not actually able to interpret obtained results. It is possible that substandard embeddings (that fail on word similarity task) correlates well with eye movement data (so this data is substandard), but it is also possible that word similarity data could be substandard itself, so eye movement data detects an actually good model.

6. Conclusions

In this paper, we compared word vectors, human judgments of word similarity and eye movement data of Russian and English languages. We noted that eye movement data is a some way correlated with word embeddings and even with word meaning, and the behavior of this data is different in other languages. Despite the results could be called negative, we can conclude that such data needs a more detailed investigation: that may be it would be more appropriate to use in another way of evaluation on eye movement data.

For example, one could try to train a regression model that gets word embeddings vectors and tries to predict one of the features of eye movement data. The set of words in eye movement data should be split into a train set and a test set, and an evaluation measure on test (for example, mean squared error) would report performance of word embeddings (the best embeddings should have the lowest error). The most reliable features could be selected by measuring p-value on predictions.

So, in future we plan to make such type of evaluation, trying to only adopt one of the features instead of their vector similarity. We also want to make a more extensive comparison obtaining other eye movement datasets for other languages (like the *Potsdam Corpus* (Stede, 2004)), and we plan to link other *neuro-cognitive resources* (like *fMRI* data) to word embeddings spaces, integrating current work in a big project about evaluation of word embeddings on different types of linguistic data.

Acknowledgements

We would like to thank three anonymous reviewers for their valuable comments and effort to improve the manuscript. We also thank my colleague, Andrey Kutuzov, for productive discussions on this paper.

7. Bibliographical References

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., and Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Auguste, J., Rey, A., and Favre, B. (2017). Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 21–26.

Baker, S., Reichart, R., and Korhonen, A. (2014). An unsupervised model for instance level subcategorization acquisition. In *EMNLP*, pages 278–289.

Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49(2014):1–47.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Fares, M., Kutuzov, A., Oepen, S., and Velldal, E. (2017). Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, number 131, pages 271–276. Linköping University Electronic Press.

Gerz, D., Vulić, I., Hill, F., Reichart, R., and Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Halawi, G., Dror, G., Gabrilovich, E., and Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

Hill, F., Reichart, R., and Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Hohenstein, S. (2013). *Eye movements and processing of semantic information in the parafovea during reading*. Ph.D. thesis, Universitätsbibliothek der Universität Potsdam.

Kennedy, A., Hill, R., and Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Kutuzov, A. and Kunilovskaya, M. (2017). Size vs. structure in training corpora for word embedding models: Araneum russicum maximum and russian national corpus. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 47–58. Springer.

Kutuzov, A. and Kuzmenko, E. (2016). Webvectors: A toolkit for building web interfaces for vector semantic models. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 155–161. Springer.

Maneewongvatana, S. and Mount, D. M. (1999). Its okay to be skinny, if your friends are fat. In *Center for Geometric Computing 4th Annual Workshop on Computational Geometry*, volume 2, pages 1–8.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Panchenko, A., Ustalov, D., Arefyev, N., Paperno, D., Konstantinova, N., Loukachevitch, N., and Biemann, C. (2016). Human and machine judgements for russian semantic relatedness. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 221–235. Springer.

Radinsky, K., Agichtein, E., Gabrilovich, E., and Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346. ACM.

Rogers, A., Drozd, A., and Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Schnabel, T., Labutov, I., Mimno, D. M., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *EMNLP*, pages 298–307.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Søgaard, A. (2016). Evaluating word embeddings with fmri and eye-tracking. *ACL 2016*, page 116.

Stede, M. (2004). The potsdam commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102. Association for Computational Linguistics.

Wehbe, L., Vaswani, A., Knight, K., and Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243.

Yang, D. and Powers, D. M. (2006). Verb similarity on the taxonomy of wordnet. In *The Third International WordNet Conference: GWC 2006*. Masaryk University.

8. Language Resource References

Cop, Uschi and Dirix, Nicolas and Drieghe, Denis and Duyck, Wouter. (2017). *Presenting GECO: An eyetrack-*

ing corpus of monolingual and bilingual sentence reading. Springer.

Laurinavichyute, AK and Sekerina, Irina Alekseevna and Alexeeva, SV and Bagdasaryan, KA. (2017). *Russian sentence corpus: Benchmark measures of eye movements in Reading in cyrillic*.

Luke, Steven G and Christianson, Kiel. (2017). *The Provo Corpus: A large eye-tracking corpus with predictability norms*. Springer.

Deep Syntactic Annotations for Broad-Coverage Psycholinguistic Modeling

Cory Shain, Marten van Schijndel, William Schuler
Ohio State, Johns Hopkins, Ohio State
Columbus, Baltimore, Columbus
shain.3@osu.edu, mvansch2@jhu.edu, schuler.77@osu.edu

Abstract

This paper presents new hand-corrected deep syntactic annotations for the sentences in two broad-coverage psycholinguistic datasets: the Dundee eye-tracking corpus (Kennedy et al., 2003) and the Natural Stories self-paced reading corpus (Futrell et al., 2017). These texts are more ecologically valid than experiment-specific constructed stimuli, allowing researchers to probe the sentence comprehension process in a naturalistic setting. Deep syntactic annotations such as categorial grammars allow direct access to phenomena like non-local or conjoined semantic argument dependencies which are relevant to many questions about sentence processing but are difficult to compute from common markup frameworks such as Penn Treebank or Universal Dependencies. Previously no gold-standard deep syntactic markups have been available for either Dundee or Natural Stories. The deep syntactic representation used for the proposed annotations (Nguyen et al., 2012) has been shown to (1) facilitate direct extraction of long-distance dependencies as well as many other syntactic constructions of interest, (2) support accurate automatic parsing, and (3) generate surprisal estimates that correlate with measures of processing difficulty (van Schijndel and Schuler, 2015). These annotations can be used for any psycholinguistic inquiry in which predictors must be computed from latent syntax trees.

Keywords: psycholinguistics, broad-coverage, treebank, categorial grammar, incremental processing

1. Introduction

Recent developments in probabilistic parsing and statistical analysis of large heterogeneous datasets have facilitated a growing interest in "broad-coverage" studies of human sentence processing in which the linguistic stimuli are rich and naturalistic rather than carefully constructed for a particular experimental purpose (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013). Instead of manipulating linguistic variables experimentally, such studies estimate measures of main effect and control variables from corpora, often using hierarchical statistical models like linear mixed effects regression (LME) (Demberg and Keller, 2008; van Schijndel et al., 2013b) or generalized additive models (GAM) (Smith and Levy, 2013) to introduce statistical rather than experimental controls.

While some linguistic variables (e.g. incremental surprisal) are best estimated in an automatic fashion using appropriate tools (van Schijndel et al., 2013a, for example), others (e.g. non-local dependency length, incremental parser operations, syntactic categories, etc.) might benefit from the use of expert syntactic annotations, which can be less noisy than automatic parses. This work presents hand-corrected deep syntactic annotations for two large broad-coverage English-language corpora: Dundee (Kennedy et al., 2003) and Natural Stories (Futrell et al., 2017). The syntactic markup used has been shown to support accurate recovery of long-distance dependencies (Nguyen et al., 2012) and to correlate with human behavior (van Schijndel and Schuler, 2015).

2. Background

2.1. Broad-coverage sentence processing research

Research into human sentence processing is concerned with understanding the mechanisms and computational procedures used by the brain to decode the linguistic signal and construct a mental representation of meaning. An important source of evidence about the structure of the human sentence processing mechanism is incremental processing effort, which can be studied using behavioral (e.g. self-paced reading, eye-tracking) or neuro-cognitive (e.g. electro/magnetoencephalography, functional magnetic resonance imaging) measures. Many theories of human sentence processing make predictions about the expected processing difficulty at a given point in an utterance as a function of syntactic features of the utterance. For example, Dependency Locality Theory (Gibson, 2000) predicts processing difficulty proportional to the length of syntactic dependencies to preceding words in the utterance. By contrast, associative memory models of sentence processing (Lewis and Vasishth, 2005; Rasmussen and Schuler, 2017) predict processing effort as a function of cue decay, which can be indexed by distance between certain decisions of a left-corner parser (Johnson-Laird, 1983).

A rich psycholinguistic literature explores theories such as these using stimuli constructed by the experimenters in order to manipulate variables of interest. For example, Grodner and Gibson (2005) manipulated dependency length by presenting subjects with sentences like

(1) The reporter who sent the photographer to the editor hoped for a story.

The use of constructed stimuli affords direct experimental control over the variable of interest as well as minimization of possible linguistic confounds. For many designs, there is also no need to model the response to every word in the utterance, only to those words that participate in critical regions as defined by the experiment (e.g. where long dependencies are resolved).

However, this experimental control of linguistic properties of the stimulus may come at the cost of introducing other confounds that might affect participants' responses. For example, the task of comprehending sentences like (1) pre-

primitive types				type-combining operators	
V	finite verb clause	**S**	top-level utterance	**-a**	argument expected ahead
I	infinitive clause	**Q**	subject-auxiliary inverted	**-b**	argument expected behind
B	base-form clause	**C**	complementized finite verb	**-c**	conjunct expected ahead
L	participial clause	**F**	complementized infinitive	**-d**	conjunct expected behind
A	adjectival/predicative clause	**E**	complementized base-form	**-g**	gap-filler
R	adverbial clause	**N**	nominal clause / noun phrase	**-h**	heavy-shift / extraposition
G	gerund clause	**D**	determiner / possessive	**-i**	interrogative pronoun
P	particle	**O**	non-possessive genitive	**-r**	relative pronoun
				-v	passive

Table 1: Nguyen et al. (2012) primitive types and type-combining operators.

sented in isolation is distinct in many ways from the usual conditions of human sentence processing. First, the words and constructions that appear in the stimuli rarely reflect the distributional characteristics of typical language use — in fact, constructed stimuli intentionally deviate from these distributional characteristics in order to test the hypothesis in question. Responses to unnatural utterances may not generalize to sentence processing in more typical cases. Second, constructed stimuli are usually presented in isolation, possibly introducing an inflated burden of pragmatic inference. For example, (1) contains three definite noun phrases, but participants are given no linguistic or situational context against which to interpret them. Third, the fact of presenting unusual sentences in isolation may signal to subjects that the implicit use of language for communication is being temporarily suspended. If it is not clear to subjects that the experimenters are trying to communicate a substantive message about the reporter, photographer, and editor, they may abandon their usual sentence processing routines and instead use task-specific heuristics. Added to the foregoing concerns about ecological validity is the fact that data collected in this way are at best difficult to reappropriate in order to study questions outside the purview of the original experimental design.

Broad-coverage studies are therefore an important complement to constructed-stimulus studies. By relaxing the requirement for direct manipulation and bringing linguistic confounds under statistical rather than experimental control, sentence processing researchers can mitigate the aforementioned problems by exposing subjects to context-rich connected texts and performing word-by-word modeling of responses to linguistic predictors computed from the stimuli. Such paradigms have been used to explore the sensitivity of the human sentence processing apparatus to variables like surprisal (Frank and Bod, 2011; Fossum and Levy, 2012; Demberg et al., 2013; van Schijndel and Schuler, 2015) and dependency locality (Demberg and Keller, 2008; Shain et al., 2016). By providing hand-corrected deep syntactic annotations for two large broad-coverage corpora, the current work aims to support further research along these lines.

2.2. Deep Syntactic Annotations

Explorations of memory effects in sentence processing typically require some indicator of precisely when during sentence processing certain syntactic arguments are attached and which semantic argument dependencies are associated with those syntactic arguments. This linguistic precision requires a deep syntactic annotation of the stimulus sentences that are the source of the modeled psycholinguistic phenomena.

The deep syntactic annotation used in this resource is a generalized categorial grammar (GCG) of English (Nguyen et al., 2012).[1] This representation both (1) defines a small set of licensed syntactic compositions, like e.g. Combinatory Categorial Grammar (Steedman, 2000), and (2) restricts the inventory of types to those needed to enforce grammatical constraints, like e.g. Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). This markup assigns a category or sign type to each meaningful sequence of words in each sentence, consisting of a *primitive clausal type* (e.g. a verb-headed clause **V**, base-form clause **B**, noun phrase or nominal clause **N**, etc.) lacking one or more dependent types, each delimited by a *type-combining operator* (e.g. **-a** to define a missing argument expected immediately ahead in the utterance, **-b** to define a missing argument expected immediately behind in the utterance, **-c** or **-d** to missing conjuncts ahead or behind, and so on). Table 1 lists the complete set of primitive types and type-combining operators in this markup. The marked up categories are constrained by a set of grammatical inference rules which assign semantic dependencies in cases of argument and modifier attachment, and keep track of these dependencies through phenomena like passive alternations, conjunctions, filler-gap constructions, extrapositions, and subject-auxiliary inversions. Table 2 lists a set of grammatical inference rules that constrain possible annotations, and Figures 1 and 2 show some example marked up sentences. This rich markup can be reliably automatically reannotated from Penn Treebank markup (Marcus et al., 1993) if available, or automatically suggested by a robust PCFG parser (Petrov and Klein, 2007, for example) trained on existing reannotated markup, with or without hand correction.

Unlike Penn Treebank markup (Marcus et al., 1993) or syntactic dependency markup (de Marneffe et al., 2006; Nivre et al., 2016, for example), unbounded dependencies are represented locally in the Nguyen et al. (2012) markup, permitting access to a store of incomplete non-local dependencies at any point in parsing. The syntactic composition rules

[1] Further in-depth details of the GCG specification are available here: `http://go.osu.edu/gcg`.

grammatical inference rules	
A	argument attachment ahead or behind
C	conjunct attachment ahead or behind
E	extraction
G	gap-filler attachment ahead or behind
H	heavy-shift / extraposition
I	interrogative clause attachment
M	modifier attachment ahead or behind
Q	subject-auxiliary inversion
R	relative clause attachment
T	type conversion / argument elision
U	auxiliary attachment ahead or behind
V	passive
X	it-extraposition
Z	zero-head introduction

Table 2: Grammatical inference rules, adapted from Nguyen et al. (2012).

map deterministically to semantic composition operations, allowing certain incremental semantic processing decisions to be recovered from syntactic annotations. The advantage of this markup for psycholinguistic modeling is the direct access it affords to incremental non-local dependency features and semantic composition operations, both of which may play a role in human sentence processing. In this respect, this markup is similar to HPSG, LFG, and various instantiations of categorial grammar such as CCG. In fact, with appropriate reannotation scripts, the present markup can in principle be used to generate these other markups automatically. GCG was selected for the present annotation because of previous work showing evidence that it has several psycholinguistically desirable properties: better automatic recovery of filler-gap and other non-local dependencies than parsers trained on dependency representations (Nguyen et al., 2012), better control over syntactic frequency confounds in psycholinguistic data than controls based on Penn Treebank annotations (van Schijndel et al., 2014), and correlation between human response times and surprisal estimates computed by an incremental parser trained on this representation (van Schijndel and Schuler, 2015).

2.3. Corpora annotated

This work presents annotations for the Dundee (Kennedy et al., 2003) and Natural Stories (Futrell et al., 2017) reading time corpora. Dundee contains eye-tracking measures from 10 subjects reading 20 editorials from *The Independent* newspaper. The stimulus set contains a total of 51,502 tokens and 2,368 sentences (Kennedy et al., 2003), with a total of 260,065 fixation events across all subjects. Dundee has been in existence for some time and has been used for psycholinguistic hypothesis testing in a variety of studies (Demberg and Keller, 2008; Frank, 2009; Frank and Bod, 2011; Fossum and Levy, 2012; Smith and Levy, 2013; Demberg et al., 2013). A treebank exists for Dundee (Barrett et al., 2015) using syntactic dependencies (Nivre et al., 2016), but syntactic dependencies are optimized for efficient parsing and as a result are not as closely related to

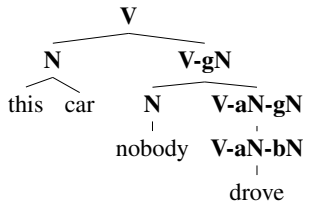

Figure 1: A simple sentence, *This car nobody drove*, annotated with Nguyen et al. (2012) markup. At the top, the noun phrase, *this car*, attaches as a gap filler (**-gN**) using inference rule G. Below that, the noun phrase, *nobody*, attaches as the first argument (**-aN**) of the verb *drove* using inference rule A. Below that, the gap filler is identified as an extracted second argument (**-bN**) of the verb *drove*, using inference rule E.

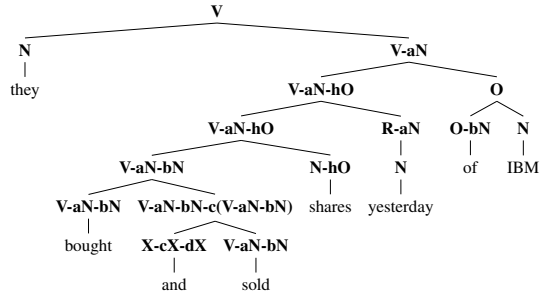

Figure 2: A more complex sentence, involving conjunction (**-c**, **-d**) of a transitive verb (**V-aN-bN**), and extraposition of a genitive complement (**-hO**) across an adverb (**R-aN**).

semantic argument structure as dependencies derived from categorial grammar markup.[2] This distinction is apparent in cases of conjunctions, extrapositions, and filler-gap extractions from embedded clauses. For example, because syntactic dependency representations typically analyze conjunctions as linked lists of conjuncts, they are not able to assign different analyses to high and low attachment readings of *old* in the conjunction *old men and women* (see Figure 3), since the word *men* serves as both the high and low site for modifier attachment. Markup based on categorial grammar or phrase structure is able to distinguish these different attachment analyses using different bracketings.

Natural Stories contains self-paced reading measures from 181 subjects reading (some subset of) 10 short stories on Amazon Mechanical Turk. The stimulus set contains a total of 10,245 tokens and 485 sentences, with a total of 848,768 reading events across all subjects. Shain et al. (2016) used Natural Stories to test hypotheses about retrieval costs during sentence processing. Natural Stories distributes with

[2]Note that there exists an enhanced deep markup for universal dependencies (Schuster and Manning, 2016) which can mitigate some of the problems with shallow dependency annotations. However, no hand-corrected deep dependency markups are available for either Dundee or Natural Stories, and many of the representational advantages of deep dependency markups are provided directly by the current GCG annotation scheme.

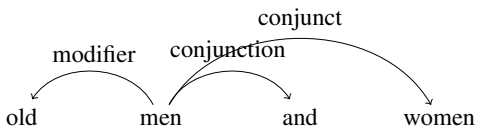

Figure 3: Syntactic dependency analysis of both high and low attachment of *old* in the conjunction *old men and women*.

hand-corrected Penn Treebank (PTB) annotations, as well as uncorrected syntactic dependency annotations automatically generated from the PTB source. The deep syntactic annotations described here complement these existing annotations by providing rich phrase-structural representations that locally encode syntactic dependency information and deterministically represent semantic composition operations, allowing researchers to more easily study the role of these features in human sentence processing.

3. Methods

Stimuli from both corpora were syntactically annotated via single-expert hand-correction of automatically-generated deep syntactic markup. In the case of Dundee, the automatic source parses were produced by the Petrov and Klein (2007) parser trained on an automatic translation from PTB to deep syntactic trees in sections 2–21 of the Wall Street Journal corpus, using the Nguyen et al. (2012) reannotation algorithm. In the case of the Natural Stories corpus, the automatic annotations were produced by applying the Nguyen et al. (2012) reannotation algorithm directly to the gold PTB-style trees supplied by the authors of the corpus (Futrell et al., 2017). Hand-correction of the Natural Stories deep syntactic reannotation was performed by a single expert annotator. The automatic parses of the Dundee corpus were partitioned in two and each set was hand-corrected by a distinct expert annotator. Depending on the complexity of the sentence, the principal annotators consulted at times with one or more additional experts before deciding on a final annotation.

4. Access

Annotations for both corpora are distributed through the ModelBlocks repository (van Schijndel and Schuler, 2013), which can be accessed at the following URL: `https://github.com/modelblocks/modelblocks-release`.[3] ModelBlocks only includes the syntactic annotations, not the stimuli themselves. Once users are in possession of the source stimuli, ModelBlocks provides scripts to automatically generate the complete treebanks by combining the annotations with the source stimuli. The Natural Stories corpus is publicly available and can be accessed at the following URL: `https://github.com/languageMIT/naturalstories`. Because of licensing restrictions on the stimuli, the Dundee corpus is not publicly available. Interested researchers must contact the authors directly (Kennedy et al., 2003) for access.

[3]This Github URL supersedes the one in the cited paper.

5. Conclusion

Because the Dundee and Natural Stories corpora are broad-coverage rather than constructed to target a particular question, they provide a more realistic measure of subjects' typical response to language stimuli, and the data they contain can be reappropriated to test a variety of hypotheses about the human sentence processing system, some of which may not have been anticipated at the time of data collection. The deep syntactic representation used here provides access to incremental non-local dependency features and semantic composition operations, which are of potential import to a range of sentence processing questions. Thus, the hand-corrected deep syntactic annotations presented in this work should have lasting value by supporting an open set of such investigations into possible determinants of sentence processing difficulty.

6. Bibliographic References

Barrett, M., Agić, Z., and Søgaard, A., (2015). *The Dundee Treebank*, pages 242–248. Association for Computational Linguistics.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demberg, V., Keller, F., and Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 39(4):1025–1066.

Fossum, V. and Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of CMCL 2012*. Association for Computational Linguistics.

Frank, S. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proc. Annual Meeting of the Cognitive Science Society*, pages 1139–1144.

Futrell, R., Gibson, E., Tily, H., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2017). The natural stories corpus. *arXiv*, (1708.05763).

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126, Cambridge, MA. MIT Press.

Grodner, D. J. and Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–291.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Kennedy, A., Pynte, J., and Hill, R. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of COLING 2012*, pages 2125–2140, Mumbai, India.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *LREC*.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411, Rochester, New York, April. Association for Computational Linguistics.

Pollard, C. and Sag, I. (1994). *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Rasmussen, N. and Schuler, W. (2017). Leftcorner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*.

Schuster, S. and Manning, C. D. (2016). Enhanced english universal dependencies: An improved representation for natural language understanding tasks. In *LREC 2016*.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 49–58. Association for Computational Linguistics.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Steedman, M. (2000). *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

van Schijndel, M. and Schuler, W. (2013). An analysis of frequency- and memory-based processing costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

van Schijndel, M. and Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.

van Schijndel, M., Exley, A., and Schuler, W. (2013a). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

van Schijndel, M., Nguyen, L., and Schuler, W. (2013b). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proc. of CMCL 2013*. Association for Computational Linguistics.

van Schijndel, M., Schuler, W., and Culicover, P. W. (2014). Frequency effects in the processing of unbounded dependencies. In *Proc. of CogSci 2014*. Cognitive Science Society.

Synesthetic Metaphors in Korean Compound Words

Charmhun Jo
Faculty of Humanities,The Hong Kong Polytechnic University
11 Yuk Choi Rd., Hung Hom, Kowloon, Hong Kong
jch337@hotmail.com

Abstract

The present study, as a follow-up research of Jo (2017), continues to test Ullmann's (1963) theoretical framework of "hierarchical distribution" through synesthetic data retrieved from Korean compound words. Namely, this study intends to judge the reliability and generalization of previous results found in synesthetic data from Korean National Corpus, and furthermore to explore the characteristics of synesthetic phenomena in compound words which have not been yet touched upon in this field. The data are gathered through the manual inspection with respect to the materials of Korean WordNet and *Standard Korean Grand Dictionary*, which are both used for the source of compound synesthesia, together with compounds from the author's intuition. As a result, Korean compound word synesthesia faithfully confirms the conclusion of Jo's (2017) study of Korean parsed corpus synesthesia which strongly supports Ullmann's (1963) synesthetic hierarchy, from the point of view of frequency tendency. The compound-word synesthesia, however, has some specificities in regards of source and target of the mappings. In other words, the role of vision is maximized as the source, while minimized as the target, and there appears no source domain in olfaction and audition.

Keywords: synesthetic metaphors, Korean compound words, hierarchical distribution, source, target

1. Introduction

In the field of linguistics, synesthesia is approached in terms of metaphor (Williams, 1976; Huang, 2015). It means that a perceptual experience of one sense is understood by lexical expressions associated with another, such as "warm color". The pioneering researcher of synesthetic metaphors is S. Ullmann (1963), who analyzed synesthetic examples from the 19th century poetic writings written in English, French, and Hungarian. Concerning the "panchronistic" natures of synesthetic mappings, Ullmann (1963) proposed his theoretical framework of "hierarchical distribution", arriving at a conclusion of three general tendencies of synesthetic transfers: firstly, the directional tendency of "touch → heat → taste → smell → sound → sight",[1] which is called "hierarchical distribution" since the transfers tend to move physically from the "lower" to the "higher" sensory domains; secondly, the source domain tendency that the most frequent source domain of transfers is touch, the lowest level of sensation; thirdly, the target domain tendency that the most frequent target domain for synesthetic transfers is sound rather than sight.

Following Ullmann's (1963) study on the synesthetic directionality, Williams (1976) investigated the synesthetic transfer patterns in ordinary language. While Ullmann's (1963) research is for synchronic data from poetry, Williams's (1976) approach focuses on diachronic data from vocabulary, namely, the historical change of meanings of synesthetic adjectives in daily English (together with some evidence from other Indo-European languages and Japanese as well). Based on his analysis of 65 English adjectives, Williams (1976) posited that the diachronic semantic change displays a highly regular movement. For instance, "dull" came out as an adjective for touch, extended to color and sound, and later to intellect or knowledge (Takada, 2008). The same pattern is also displayed in other Indo-European languages and Japanese. In summary, the findings of Williams (1976) on synesthetic metaphors in ordinary language support Ullmann's (1963) framework of "hierarchical distribution".

Day (1996) examined synesthetic occurrences collected from the printed and electronic texts of English, and proposed a "general distribution" of synesthetic metaphors, as shown in the following: touch → taste → temperature → smell → sound → sight. It signifies that the synesthetic metaphor transfers at large go from the "lower" to the "higher" sensory modes in the same manner as the findings of Ullmann (1963) and Williams (1976). In the meanwhile, Shen (1997), in terms of cognitive poetics, explored the directional tendency of mapping for Hebrew synesthesia based on the literary analysis of modern poetry and two psycho-linguistic experimental data. His results strongly confirmed Ullmann's (1963) observation about the synesthetic hierarchy. That is to say, the synesthetic expressions in Hebrew tended to map lower perceptions on to higher ones in their hierarchical order. Via the notion of "accessibility", Shen (1997) suggested that the "low to high" transfer comes from the general cognitive constraints where "a mapping from more 'accessible' or 'basic' concepts onto 'less accessible' or 'less basic' ones seems more natural, and is preferred over the opposite mapping". He also pointed out that sight and sound are less accessible because they do not involve any direct contact with the perceived entity. To verify the "universal" validity of the synesthetic hypothesis claimed by Ullmann (1963) and Williams (1976), Yu (2003) analyzed synesthetic data extracted from literary works written by current Chinese writer, Mo Yan, based on a "cognitive perspective". The results of the research demonstrated that Chinese synesthesia basically complies with their general schemes in metaphoric mappings as well.

[1] This sign "A → B" signifies that A (the source) is mapped onto B (the target) between sensory domains, A modifying B. In the study of Ullmann (1963), the term "mapping" is not used, but instead he uses the term "transfer". Also, the term of "target" do not appear in the original report, but instead "destination" or "recipient" is employed. Additionally, concerning the sensory domains utilized, Ullmann (1963) selected six senses including "heat" separated from "touch", as seen in the above. That is why some scholars simplify his hierarchy into "touch → taste → smell → sound → sight".

Until now, the linguistic subjects examined for synesthetic phenomenon have been steadily expanded from English to other languages such as Italian, Hebrew, and Chinese, as Ullmann (1963) and Williams (1976) presenting probable universal principles in the process of synesthetic association both require broader investigations of more linguistic samples so that their theories can be built up universally. Despite that, many languages, including Korean, have been still remaining to be dealt with. In this respect, the present study reported in this article, as a follow-up research of Jo (2017), continues to test Ullmann's (1963) theoretical framework of "hierarchical distribution" through the synesthetic data retrieved from Korean compound words. In other words, focusing on the issue of the directionality of linguistic synesthesia rather than that of its motivation, this study intends to judge the reliability and generalization of the previous results found out in the synesthetic data coming from Korean National Corpus (KNC), and furthermore to explore the characteristics of synesthetic phenomena in compound words which have not been yet touched upon in this field.[2]

In what follows, this paper presents a brief literature review of the tendencies of synesthetic mappings in Korean ordinary language reported in Jo (2017) in the second section. The research methods including data collection are then presented in the third section, and the results analyzed are laid out in the fourth section, followed by a general discussion. In the last section, the conclusion of the current study is given along with a summary.

2. Literature Review: Synesthesia in Korean Ordinary Language

Several research works that have addressed Korean synesthetic phenomena so far based on Ullmann (1963) or Williams (1976), have not yet showed a certain clear and comprehensive directional order of synesthetic transfers or their obvious findings regarding that (e.g., Yoon, 1970; Park, 1978 for Korean poetic synesthesia, and Chung, 1997; Lee, 2015 for Korean daily language synesthesia). In this situation, Jo (2017) attempted to clarify the regularities and features of Korean synesthesia based on the clear-cut data via the corpus-based approach. Exactly, he investigated synesthetic data extracted from the KNC parsed corpus[3] and compared the findings with those from Ullmann (1963). The overall result of synesthesia collected from the Korean parsed corpus is arranged below in Table 1. It demonstrates an overview of corpus work upon Korean synesthetic occurrences.

Total Corpus Sentences (TCS)	Extracted Positive Sentences (EPS)	True Positives (real synesthesiae) (TP)	TP / EPS (%)	TP / TCS (%)
43,828	1,250	100	8	0.23

Table 1. Overall synesthetic transfer route in KNC, proposed by Jo (2017)[4]

Below is the overall distribution of synesthetic mappings among sensory modes in KNC. This data is substantial information on Korean conventional synesthesia retrieved from corpus.

Target / Source	Touch	Taste	Smell	Sight	Hearing	Total
Touch	0	3	3	11	20	37
Taste	1	0	8	9	15	33
Smell	0	0	0	1	2	3
Sight	2	1	4	0	13	20
Hearing	0	1	1	5	0	7
Total	3	5	16	26	50	100

Table 2. The distribution of synesthetic mappings among sensory domains in KNC (TOKEN), presented by Jo (2017)

Accordingly, from the synesthetic data presented in Table 2, Jo (2017) set up the overall synesthetic transfer route in Korean ordinary language as follows:

Touch → Taste → Smell → Sight → Hearing

Figure 1. Overall synesthetic transfer route in KNC, proposed by Jo (2017)

Of course, this is based on the frequency of mappings, according to which the "forward" tokens account for 85% and the "backward" ones just account for 15%. Based on these results, the researcher suggested conclusively that the directional order of Korean synesthesia generally corresponds to the directions from Ullmann (1963) and Williams (1976), and that the most frequent source and target domains of the synesthetic transfers investigated are in accordance with Ullmann's (1963) findings, as touch and

[2] To the author's knowledge, there are no previous studies upon compound words with respect to synesthesia yet.

[3] The study by Jo (2017) basically followed Strik Lievers et al.'s (2013) methods to extract synesthetic data from KNC. The way can be summarized as follows: firstly, for the sense-related word lists, the lexical items are compiled, subdivided by five sensory domains respectively in terms of POS categorization of verb (V), adjective (A), and noun (N), which start from the intuition and the relevant literature and are expanded via some available electronic resources such as Korean WordNet and web dictionaries in KNC; secondly, as for the synesthesia extraction from the corpus, a simplest method that just lists all the sentences containing at least two perception-related words is applied to this KNC parsed corpus unlike Strik Lievers et al.'s (2013) methodology, given the fact that this simplest way can possibly collect the largest number of candidate sentences and the candidates will be affordable for the final manual checking because the corpus is not big relatively; lastly, to sort out "true" synesthesiae, it is necessary to do a hand work inspection of the extracted candidate output.

[4] Offering further elaborations, "TP" means 100 tokens of synesthetic occurrences finally detected from "EPS", i.e., 1,250 candidate sentences where true synesthesia could be found. As obviously recognized in the percentages of "TP/EPS" and "TP/TCS", the rarity of synesthesia in quantity in ordinary language is verified, although it is very common in daily use.

sound each.[5] Also, Jo (2017) pointed out that there can exist a "delicate cultural dependency" with regard to Korean synesthesia, interpreting that the difference of the proportion between the most and second frequent source sensory domains is very slight as the following:
"This situation can imply that together with the tactile domain, touch, the sense of taste takes up a significant position in Korean or Asian cultural context, and so people in the cultural circle more often tend to describe something in terms of gustation or tactility, compared with western people."

As aforementioned in relation to Korean synesthesia, Jo's (2017) study is probably the "first" attempt for the setup of the directionality of Korean synesthetic mappings based on an obvious and extensive database. Thus, his model for Korean synesthetic metaphors still needs to be confirmed by subsequent examinations with another synesthetic data

3. Methodology

3.1 Sensory Domains

There is no agreement among scholars over how many sensory modalities there exist, and they can vary depending upon the researchers' perspective and classificatory criteria (Strik Lievers et al., 2013; Strik Lievers, 2015). Most of synesthetic studies now follow the Aristotelian five-sense system of touch, taste, smell, sight, and hearing (cf. Cytowic, 1989; Shen, 1997; Strik Lievers, 2015).

The study reported in this paper selects the general Aristotelian five sensory modes for the harmonious comparison with the results from Jo (2017). The details including sensory domains and organs are displayed below:

Sensory domain	Sub-categorical sensory mode	Sensory organ	Sensory object
Touch	contact, temperature/heat, pain, hardness, tightness, humidity, texture, pressure, etc.	hands and skin	physical and non-physical entities (e.g., toys, water, wind)
Taste	sweetness, saltiness, spiciness, sourness, bitterness, etc.	tongue	physical entities (e.g., food, drinks)
Smell	quality, quantity, intensity, etc.	nose	smell and fragrance
Sight	dimension (size, length, height, width, depth, thickness, etc.), color, form/shape, appearance, etc.	eyes	visible entities (e.g., buildings, clouds, sky, smoke, rainbows)
Hearing	quality, quantity, intensity, etc.	ears	sound and voice

Table 3. Five sensory domains and relevant information

3.2 Taxonomy

In order to facilitate the understanding and convenience of analysis of synesthetic expressions in linguistics, it is necessary to try to internally classify their types in brief. In terms of formational structure, synesthetic metaphor can be divided into three types such as lexical level synesthesia, phrasal level synesthesia, and sentential/syntactic level synesthesia.[6] At the lexical level of linguistic synesthesia, again, there can be two sub-types, namely, single word synesthesia and compound word synesthesia.[7]

The previous study by Jo (2017) for Korean conventional synesthesia focuses on the phrasal and syntactical synesthetic instances, whereas the present study deals with synesthetic examples from compound words at the lexical level.

3.3 Data

For Korean compound word synesthesia, the data will be gathered through the manual inspection with respect to the materials of Korean WordNet [8] and *Standard Korean Grand Dictionary*[9], which are both used for the source of compound synesthesia, together with compounds from the author's intuition.[10] Due to the time limit, the current study could not exceed 50 instances. Williams's (1976) diachronic study on lexical level synesthesia is based on the examination of 65 adjectives, making use of *Oxford English Dictionary* and *Middle English Dictionary*.

This exploration upon compound word synesthetic metaphors might contribute to developing a new significant research issue in the field of lexical semantics as well as expanding the research area of linguistic synesthesia.

[5] As shown in Table 2, tactition is the source in 37%, and audition is the target in 50%.

[6] Examples of phrasal and syntactic synesthesia: "cold color", "warm words"; "It smells salty", "It sounds sweet". These types are in general known as the structurally most common synesthetic metaphors.

[7] For examples of single word synesthesia, refer to Williams's (1976) survey on the historical semantic shift of English adjectives from one sensory modality to another.

[8] Access: http://www.wordnet.co.kr/. For further information with reference to Korean WordNet, refer to Chagnaa et al. (2007), Choi and Kim (2008), or Moon (2010) among others.

[9] Access: http://stdweb2.korean.go.kr/main.jsp.

[10] The analyses of compound words in this study are mainly based on compound verbs combining with auxiliary verbs such as *tay-ta* (touch) or *po-ta* (see). Although it is widely accepted that theses auxiliary verbs already went through grammaticalization (Sohn, 2001), such cases are also considered as synesthesia here in this study in terms of examples such as "noisy color". Yoon (1970) mentioned them as synesthetic phenomena in his research as well.

4. Results and Discussion

4.1 Results

The total number of the compound-word synesthesiae found is forty-five (tokens), with forty-three types.[11] The overall distribution of synesthetic mappings among sensory modes in Korean compound words is illustrated below in Table 4:

Target / Source	Touch	Taste	Smell	Sight	Hearing	Total
Touch	0	3	2	1	8	14
Taste	0	0	5	1	7	13
Smell	0	0	0	0	0	0
Sight	2	5	1	0	10	18
Hearing	0	0	0	0	0	0
Total	2	8	8	2	25	45

Table 4. The distribution of synesthetic mappings among sensory domains in Korean compound words (TOKEN)

As showed in Table 4, in the transfers of synesthetic phenomena in Korean compound words, the predominant sensory source mode is sight and the predominant target is hearing. More precisely, the visual domain acts most frequently as the source in 18 of the 45 collected synesthesiae, followed by the tactile domain in 14, while the auditory domain becomes the largest target in 25, followed by the gustatory and olfactory domain in the same number of 8 respectively.

The representative examples of Korean compound-word synesthesia are as follows[12]:

(1) Touch → Taste
먹어대다
mek-e-tay-ta
eat-P-touch-P
'keep eating'

(2) Touch → Smell
맡아대다
math-a-tay-ta
sniff-P-touch-P
'keep sniffing'

(3) Touch → Sight
쏘아대다
sso-a-tay-ta
glower-P-touch-P
'keep glowering (at someone)'

(4) Touch → Hearing
외쳐대다
oychi-e-tay-ta
shout-P-touch-P
'keep shouting'

(5) Taste → Smell
a. 쓴내
ssu-n-nay
bitter-P-smell
'bitter smell'
b. 단내
ta-n-nay
sweet-P-smell
'sweet smell'

(6) Taste → Sight
쓴웃음
ssu-n-wusum
bitter-P-smile
'wry smile'

(7) Taste → Hearing
쓴소리
ssu-n-soli
bitter-P-sound
'criticism (or bitter remark)'

(8) Sight → Taste
먹어보다
mek-e-po-ta
eat-P-see-P
'try eating'

(9) Sight → Hearing
a. 가려듣다
kali-e-tut-ta
select-P-listen-P
'listen selectively'
b. 잔소리
ca-n-soli
small-P-sound
'nagging (or nagging voice)'

4.2 General Discussion

Based on the synesthetic data reported in Table 4, the linear model for synesthetic associations in Korean compound words can be displayed as the following:

Touch → Taste → Smell → Sight → Hearing

Figure 2. Overall synesthetic transfer route in Korean compound words

In the above frequency-based model, the mappings in the direction of the arrow occupy approximately 82%, while those in the counter direction of the arrow take up approximately 18% of the total mappings. The proportions are similar to those of the earlier synesthetic data from KNC parsed corpus. In the synesthetic metaphors from compound words, there is no case transferring to all other

[11] See Appendix for the entire synesthetic expressions from Korean compound words.

[12] In this paper, each Korean language example will be described at four levels: first, *Hangul* as Korean writing system, second, phonetic transcription by *Yale Romanization*, third, gloss literally in English, and fourth, English translation. In addition, the notation for gloss in lexical analysis is simplified with P as particle.

domains from smell and sound, as showed in Table 4. The illustration in Figure 2, hence, can be again described fine-tuned below:

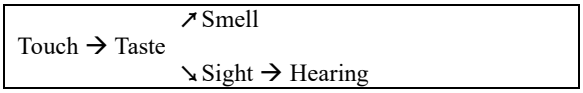

Figure 3. Overall synesthetic transfer route in Korean compound words, re-adjusted

The above directional tendency is in line with the results of Ullmann (1963) and Williams (1976). In particular, it is more similar to that of Williams (1976), given that dimension and color in his adopted sensory domains are combined together into vision. For their comparison, below is the synesthetic hierarchy proposed by Williams (1976):

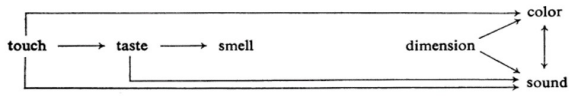

Figure 4. Synesthetic transfer route of Williams (1976)

In this respect, the directionality of the Korean conventional synesthesia including both corpus and compound word synesthetic data conforms to the Ullmann's (1963) theoretical framework of "hierarchical distribution".

However, with regard to the frequency of the source and target of the mappings, the compound-word synesthesia shows somewhat a different aspect to Ullmann's (1963) hypothesis. That is, the largest source here is not touch (about 31%) but sight (40%), which does not match with Ullmann's (1963), and the largest target is sound (about 56%), which matches with his theory. Specifically, the distributions of the source and target sensory domains in synesthetic mappings from Korean compound words are summarized below.

Sight	Touch	Taste	Hearing	Smell
40	31	29	0	0

Table 5. Source sensory domains in frequency-decreasing ordering in synesthetic mappings from Korean compound words (%, approximately)

Hearing	Smell	Taste	Sight	Touch
56	18	18	4	4

Table 6. Target sensory domains in frequency-decreasing ordering in synesthetic mappings from Korean compound words (%, approximately)

Here, it is necessary to compare two above results to the corresponding ones of KNC synesthesia in Jo (2017) and English/Italian synesthesia in Strik Lievers (2015):

Touch	Taste	Sight	Hearing	Smell
37	33	20	7	3

Table 7. Source sensory domains in frequency-decreasing ordering in synesthetic mappings from KNC (%), presented by Jo (2017)

	Touch	Taste	Sight	Hearing	Smell
English	49.3	25.7	21.8	3.0	0.2
Italian	55.6	20.2	19.1	4.6	0.2

Table 8. English and Italian source sensory domains in frequency-decreasing ordering (%), adapted from Strik Lievers (2015)

	Hearing	Sight	Smell	Taste	Touch
Korean	50	26	16	5	3
English	52.3	28.0	12.4	5.3	2.1
Italian	50.2	42.5	3.8	3.0	0.2

Table 9. Target sensory domains in frequency-decreasing ordering in Korean, English, and Italian (%), adapted from the data presented in Strik Lievers (2015) and Jo (2017)

From the above, the visual modality in Korean compound-word synesthetic metaphors is certainly noticeable in the midst of the situation following the "general" ordering by and large. In other words, the role of sight is maximized as the source, while minimized as the target. Also, we can recognize that touch and taste play a considerable role as the source, as displayed in Table 6. On top of that, another noticeable point here is that there is no source domain in smell and hearing, which can provide an interesting research topic in relation to the understanding of the cause, e.g., whether its cause is connected to the matter of the grammatical and combinational structure emerging from the lexical level of synesthesia. It is hard to jump to a conclusion from the above data yet for now, and so they need to wait for follow-up studies via more synesthetic examples.

5. Conclusion

From the above discussion upon the directional tendency of Korean synesthesia, in sum, the result from Korean compound word synesthetic data faithfully confirms the conclusion of Jo's (2017) study of Korean parsed corpus synesthesia which strongly supports Ullmann's (1963) synesthetic hierarchy, from the point of view of frequency tendency with no absolute restriction such as uni-directionality. The compound-word synesthesia, however, has some specificities in regards of source and target of the mappings. Namely, the role of vision is maximized as the source, while minimized as the target, and there appears no source domain in olfaction and audition.

For future works, accordingly, additional investigations into Korean compound word synesthesia are required with more synesthetic data in order to clarify the unique features. Also, to further affirm the tendencies of Korean synesthesia, the research of synesthetic data from Korean poetry should be conducted, given that Ullmann's

(1963) "universal" hypotheses emerged from a series of explorations into poetic language. Additionally, the issue of motivation with regard to synesthetic metaphor remains to be addressed, in particular, in terms of providing a bridge between neuro-scientific approach to synesthesia and linguistic approach to synesthesia.[13]

6. Acknowledgements

This research came from a part of the author's doctoral thesis to be submitted to The Hong Kong Polytechnic University. I wish to take an opportunity here to express my deep gratitude and respect to my supervisor Dr. Chu-Ren Huang and co-supervisor Dr. Sun-A Kim, for their invaluable advice and feedback together with constant encouragement throughout the study. Their valuable suggestions and comments were of huge assistance for my research work.

7. Appendix

Korean synesthetic expressions from compound words (TOKEN)

TOUCH → TASTE
1. 먹어대다 Keep eating
2. 마셔대다 Keep drinking
3. 빨아대다 Keep sipping

TOUCH → SMELL
4. (냄새 등) 맡아대다 Keep sniffing
5. 풍겨대다 Keep giving off (odor)

TOUCH → SIGHT
6. (시선 등) 쏘아대다 Keep glowering (at someone)

TOUCH → HEARING
7. (음악 등) 틀어대다 Keep playing (music)
8. 외쳐대다 Keep shouting
9. 불러대다 Keep singing (or calling)
10. (악기 등) 불어대다 Keep blowing (wind instruments)
11. 외워대다 Keep reading out loud (to memorize)
12. 읊어대다 Keep reciting
13. 소리치다 Yell
14. 고함치다 Shout

TASTE → SMELL
15. 쓴내 Bitter smell
16. 단내 Sweet smell
17. 쉰내 Rancid smell
18. 짠내 Salty smell
19. 비린내 Fishy smell

TASTE → SIGHT
20. 쓴웃음 Smirk (or wry/bitter smile)

TASTE → HEARING
21. 귀먹다 Deaf
22. 귀머거리The deaf
23. 쓴소리 Criticism (or bitter remark)
24. 쉰소리Hoarse sound
25. 쉰목소리 Hoarse voice
26. 고언 (苦言) Exhortation (or pungent remark)
27. 감언 (甘言) Flattery (or sweet talk)

SIGHT → TOUCH
28. 만져보다 Try touching (see how it feels)
29. 대보다 Try touching (see how it feels or how it measures)

SIGHT → TASTE
30. 먹어보다 Try eating
31. 마셔보다 Try drinking
32. 빨아보다 Try sipping
33. 맛보다 Try tasting
34. 맛보기Tasting

SIGHT → SMELL
35. (냄새 등) 맡아보다 Try sniffing

SIGHT → HEARING
36. (소리 등) 들어보다 Try listening (to sounds)
37. 외쳐보다 Try yelling
38. 읊어보다 Try reciting
39. 외워보다 Try memorizing
40. 소리쳐보다 Try shouting
41. 불러보다 Try calling
42. (악기 등) 불어보다 Try blowing (wind instruments)
43. 새겨듣다 Listen carefully
44. 가려듣다 Listen selectively
45. 잔소리 nagging (or nagging voice)

8. Bibliographical References

Chagnaa, A., et al. (2007). On the Evaluation of Korean WordNet. *Lecture Notes in Computer Science, 4629*.

Choi, U., & Kim, K. (2008). Development of Korean WordNet and Thesaurus. *Korean Lexicography, 11*. [Korean]

Chung, I. (1997). Synaesthetic transfers of Korean adjectives. *Studies in Modern Grammatical Theories, 11*, 163-180. [Korean]

Cytowic, R. E. (1989). *Synesthesia: A Union of the Senses.* New York: Springer-Verlag.

Day, S. (1996). Synaesthesia and synaesthetic metaphors. *Psyche*, *2*(32), 1-16.

Huang, C. R. (2015). Towards a lexical semantic theory of synaesthesia in Chinese. *Keynote Speech in the 16th Chinese Lexical Semantics Workshop (CLSW-16)*. Beijing.

Huang, C. R. (2016). Synaesthesia: Language, Thought, Cognition and Culture. *IEICE technical report*, *116*(368), 111-113.

Jo, C. (2017). A corpus-based study on synesthesia in Korean ordinary language. *PACLIC31*. Retrieved from http://paclic31.national-u.edu.ph/wp-content/uploads/2017/11/PACLIC_31_paper_72.pdf.

Lee, S. (2015). A study on synesthetic metaphors in Korean and Chinese advertisements from the viewpoint of cognitive linguistics. *Chinese Literature, 68*, 205-242. [Korean]

[13] "Neurological studies focus on synaesthesia as a special neuro-cognitive condition while linguistic studies focus on synaesthesia as conventionalized linguistic device. Hence studies of synaesthesia in these two fields rarely overlap. There is an urgent need to provide a bridge between these two approaches." (Huang, 2016, p. 111)

Moon, K. (2010). Criteria for Sub-category of Korean "human" Lexical field. *New Korean Education, 85*. [Korean]
Park, G., (1978). Synesthetic mappings of poetic diction. *Korean Language and Literature*, No.78. [Korean]
Shen, Y. (1997). Cognitive constraints on poetic figures. *Cognitive Linguistics*, *8*(1), 33-72.
Sohn, H. (2001). *The Korean Language*. Cambridge: Cambridge University Press.
Strik Lievers, F. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of Language*, *22*(1), 69-95.
Strik Lievers, F., et al. (2013). A methodology for the extraction of lexicalized synaesthesia from corpora. *ICL (International Congress of Linguists) 19*. Geneva, Switzerland.
Ullmann, S. (1963). *The Principles of Semantics* (2nd ed., 3rd Impression). Oxford: Basil Blackwell.
Yoon, H. (1970). The structurality of synesthetic metaphors. *Korean Language and Literature*, No.49-50. [Korean]
Yu, N. (2003). Synesthetic metaphor: A cognitive perspective. *Journal of Literary Semantics*, *32*(1), 19-34.
Williams, J. M. (1976). Synesthetic Adjectives: A possible law of semantic change. *Language*, *52*, 461-478.

Frequency and Predictability Effects in Natural Reading: Evidence from Co-Registration of Eye-Movement and Event-Related Potentials Measures

Chun-hsien Hsu
Institute of Linguistics, Academia Sinica, Taiwan
kevinhsu@gate.sinica.edu.tw

Chia-ying Lee
Institute of Linguistics, Academia Sinica, Taiwan
chiaying@gate.sinica.edu.tw

Jie-li Tsai
Department of Psychology, National Chengchi University, Taiwan
jltsai@nccu.edu.tw

Keywords: eye-movement, ERP, word frequency, predictability

1. Introduction

Language comprehension involves retrieval of individual words embedded in sentences (bottom-up processing) and the contextual-based prediction for the upcoming words (top-down processing). Word frequency and predictability are well-established effects to reflect the bottom-up and top-down processes, respectively. Studies of eye-movement in natural reading have consistently reported longer fixation and gaze durations (GD) for reading low frequency words than high frequency words (Kliegl et al., 2006). Meanwhile, GD on words which are predictable based on the preceding context are shorter than that on when words which are unpredictable. These eye movement measures indicate that contextual information plays a role for word processing and text integration. Event-related potential (ERP) studies have also demonstrated that N400, an ERP component to index the lexical retrieval and semantic integration, are inversely proportional to the cloze probability and word frequency (Kutas and Hillyard, 1980). However, in the traditional ERP studies, words were presented serially at the fixed rate, the so-called serial visual presentation. Although this procedure allows researchers to avoid the contamination from the saccade-related potentials in EEG recording, it is unclear whether the ERPs findings could be applied to natural reading. This study aims to address this issue by simultaneously record eye movement and ERP data from 27 participants in reading 160 sentences of the Taipei Sentence Corpus (TSC) to examine how word frequency and word predictability play roles in natural reading. The co-registration of eye-movement and ERPs data may provide new perspectives to examine the relationships among lexical properties and single-trial ERP measurements.

2. Design and Material

We utilized the Taipei Sentence Corpus (TSC) (Tsai, 2013) that comprises 160 Chinese sentences and each sentence contains 20 to 26 characters (M = 22.3, SD = 1.7). Word frequency, word length, and word strokes of all words in the sentences were acquired from Sinica Corpus 4.0. The range of word length in the TSC was from 1 to 4 characters (M = 1.58, SD = 0.6). There were 2246 words in total and the number of words for word length from 1 to 4 are 1025, 1151, 49, and 21 respectively.

3. Apparatus and Data Acquisition

Eye movements are recorded via an SR Research Eyelink 1000 long-range MRI-compatible eye tracker with a sampling rate of 1000 Hz. The size of a character presented on the screen was 32 x 32 pixels, and there was a space of 4 pixels between characters. The continuous EEG was recorded from 64 Ag/AgCl electrodes (QuickCap, Neuromedical Supplies, Sterling, USA) digitized at a rate of 1000 Hz using a SynAmps2 ® (Neuroscan, Inc.) amplifiers. For offline analysis, a 700 ms segment of EEG was cut (from 100 ms before to 600 ms after fixation onset) for each fixation. Because the signal-to-noise ratio of EEG is poor, we applied a novel approach described in Tzeng et al. (2017) to measure N400 in single-trial data. First, the ensemble empirical mode decomposition (EEMD) approach was used to extract event-related modes (ERMs) that reflect 4-8Hz EEG oscillations. Than, mean amplitudes of N400 were measured in the 300–350 ms interval after onsets of fixations in channel P4. Statistical analysis was performed using the linear-mixed model (LMM) including participants, words and sentences as random effects, and word frequency, predictability, word position, and interactions as fixed effects.

4. Results and Conclusions

While using GD as the dependent variable, the LMM model revealed significant effects of word frequency (*beta* = -.056, *S.E.* = .003, t = -17.68, p < .001), prediactiability (*beta* = -.029, *S.E.* = .004, t = -7.21, p < .001), and frequency by predictability interaction (*beta* = .01, *S.E.* = .002, t = 4.85, p < .001). For N400, the LMM model also revealed a significant effect of word frequency by predictability interaction on N400 (*beta* = -.168, *S.E.* = .084, t = -2.001, p < .05)

In summary, the presented study used fixation-related ERPs to evaluate word frequency and predictability effects on GD and N400 during natural reading. The word frequency by word predictability interactive effect on N400 and GD reflects that low predictability words elicited a larger N400 and longer GD than high predictability words, and the predictability effect on low frequency words is stronger than that on high frequency words. These results support that fixation-related ERM is sensitive enough to reveal effects of contextual-baed prediction and lexical retrieval during natural reading.

5. Bibliographical References

Kliegl, R. et al. (2006). *Journal of experimental psychology: General*, 135:12.

Kutas, M., & Hillyard, S. A. (1980). *Science*, 207:203-205.

Tzeng, Y.L. et al. (2017). *Frontiers in psychology*, 8:433.

Tsai, J. L. (2013). Technical report of National Science Council, NSC 99-2410-H-004-091-MY2.

Reduced Syntactic Processing Efficiency in Older Adults in Reading Sentences

Zude Zhu, Xiaopu Hou, Yiming Yang
School of Linguistic Sciences and Arts, Jiangsu Normal University, Xuzhou China 221009
Collaborative Innovation Center for Language Competence, Xuzhou China 221009
Jiangsu Key Laboratory of Language and Cognitive Neuroscience, Xuzhou China 221009
Institute of Linguistic Sciences, Jiangsu Normal University, Xuzhou China 221009
zhuzude@163.com, xiaopu_hou@qq.com, yangym@jsnu.edu.cn

Keywords: aging, P600, syntactic processing, efficiency

1. Introduction

Sentence comprehension is one of the key components of human language. In order to construct a meaningful representation of a given sentence, one has to recognize single words and integrate word-level semantic blocks into a larger semantic utterance under the guidelines of semantic and syntactic rules. While research has documented this age-related decline in semantic processing, there is still debate concerning whether syntactic processing also declines during aging. The aim of the present study was to test whether syntactic processing, in addition to semantic processing, declines during aging.

2. Methods

To control for the confounding effects of other cognitive skills, the recruited 26 younger and 20 older adults were well matched on working memory capacity, general intelligence, verbal intelligence, and verbal fluency. The study included congruent sentences (CON, 妹妹把窗户**擦洗**完了/ The younger sister **cleaned up** the window.), sentences with semantic violation (SEM, 妹妹把窗户**抄袭**了/ The younger sister **plagiarized** the window.), and sentences with both semantic and syntactic violation (SEM+SYN, 妹妹把窗户**茶叶**了/ The younger sister **tea** the window.) (Wang et al., 2008). The sentences differ from each other only on the critical words, which were matched for word frequency and number of strokes. The three conditions showed significant differences in terms of semantic acceptability. The two incongruent conditions were both rated as significantly less acceptable than the congruent condition, and there was no significant difference between the two incongruent conditions. The SEM vs. CON and SEM+SYN vs. SEM contrasts would reveal semantic and syntactic effect, respectively. EEG data were recorded with 64 channel Neuroscan system. After preprocessing, ground averaged event-related potential (ERP) data were used for comparison.

3. Results and Discussion

The behavioral results revealed that the older adults had significantly lower accuracy on measures of semantic and syntactic processing compared to younger adults. For ERP data, the older adults showed delayed peak latency of N400 and P600 compared to the younger adults for semantic analysis. For the N400 amplitude, there was a significant Condition by Region by Hemisphere by Group interaction, simple main effects revealed significantly higher N400 amplitude in the SEM condition compared to the CON condition in anterior and posterior regions in both the left and right hemisphere for younger adults, whereas the same was true only in the right hemisphere for older adults.

In the syntactic analysis, the older adults showed delayed peak latency of N400 and P600 compared to the younger adults. For N400 amplitude, there was a significant Condition by Hemisphere by Group interaction. Simple main effect revealed that the N400 effect was found in both anterior and posterior regions and in the left and right hemispheres for younger adults, and was found in the posterior region but not in the anterior region in the older adults. For the syntactic processing related P600 amplitude, there was no significant Group by Condition interaction in either the left anterior, left posterior, right anterior or right posterior regions. Critically, a larger P600 effect was associated with lower accuracy in the SEM+SYN condition compared to the SEM condition for the older adults but not for the younger adults.

While the P600 effect suggests that the older adults were able to respond to the syntactic violation as younger adults did (Kemmer et al., 2004), the key finding of the present study was that syntactic processing was less efficient in older adults. During syntactic processing, older adults also showed delayed peak latency for P600 relative to younger adults. Moreover, behavior-ERP correlation analysis revealed that the larger P600 effect was linked with less accuracy in the SEM+SYN condition compared to the SEM condition in the older group only. The correlation results suggested that the P600 effect in the older adults reflected a less efficient response (Zhu et al., 2015) indicating they failed to effectively use syntactic information during reading. In summary, the present study is the first to document that syntactic processing declines during aging in addition of semantic processing decline.

4. Acknowledgements

The work was supported by grants from the Natural Science Foundation of China (NSFC 31571156) and the National Social Science Foundation of China (17ZDA301).

5. Bibliographical References

Kemmer, L., Coulson, S., De Ochoa, E., & Kutas, M. (2004). Syntactic processing with aging: an event-related potential study. *Psychophysiology*, 41, 372-384.

Wang, S., Zhu, Z., Zhang, J.X., Wang, Z., Xiao, Z., Xiang, H., & Chen, H.C. (2008). Broca's area plays a role in syntactic processing during Chinese reading comprehension. *Neuropsychologia*, 46, 1371-1378.

Zhu, Z., Johnson, N.F., Kim, C., & Gold, B.T. (2015). Reduced frontal cortex efficiency is associated with lower white matter integrity in aging. *Cerebral Cortex*, 25, 138-146.

The Stroop-like Effect During Sound Perception Task in Bilingual Minds

Libo Geng[1] , Lillian Zhao[2] , Jiaoyan Fang[1]

[1] School of Linguistic Science and Arts, Jiangsu Normal University1, Heping Rd. 57, Xuzhou, Jiangsu Province, China

[2] School of Science, Dartmouth College, Hanover, NH 03755 USA

genglb@jsnu.edu.cn, Lillianzhao@gmail.com, fangjiaoyan@foxmail.com

Keywords: Stroop-like Effect, ERP, Perception Task, Bilinguals

1. Introduction

It is widely accepted that concepts can be represented in two linguistic forms in the bilingual lexicon: first language (L1) and second language (L2). Recently, whether the phonology in the L1 interferes with the word recognition in the L2 for bilingual people is a subject of lively debate.

There is an assumption that there may be a kind of Stroop-like Effect, which was first proposed in psychology, in bilinguals' brain. When bilinguals read or listen L2 word pairs(WP) with a sound-similarity judgment task and if L1 sound is also activated, there will be 4 conditions as shown in Table 1.

Based on this assumption, the paper will examine this issue in sound similarity judgment tasks by assessing effects of phonological interface on the recognition of Chinese disyllabic word pairs by English-Chinese bilingual speakers. What the paper hypothesis is there is a kind of Stroop-like effect in bilingual brains when the result of explicit processing of L2 word pairs is not the same as the implicit processing of L1 word pairs.

2. Background

Different with the cross-language tasks, which encourages a bilingual activation pattern, in one-language tasks cross-language effects of L2 on L1 were found in a purely L1 context (Kroll and Stewart, 1994；Hell and Dijkstra, 2002). Thierry and Wu (2007) found bilinguals' knowledge of their L1 was activated in the context of their L2. Participants exhibited a reduced N400 amplitude for English word pairs whose Chinese translations shared a common character relative to English word pairs whose Chinese translations did not share a common character, the temporal parameters and necessary characteristics of the task may determine whether bilinguals access the L1. (Van Heuven et al. 2008). Obviously, the research is a good support to unconscious activation of L1 during L2 semantic comprehension task. What will be the sound-perceptive task? When bilinguals just read or listen to words in their L2 with perceptive task, will the L1 be activated?

3. Experiment Methodology

The electrophysiological experiment was carried out as below. 20 English-Chinese (Mandarin) bilinguals whose L1/L2 is in different levels and 20 monolingual Chinese (Mandarin) speakers were selected as participants in our researches to control any potential priming effects. Then the 200 word pairs as shown in Table 1 used were matched across experimental conditions for lexical frequency and word concreteness.

WP L1-L1 WP L2-L2	Common first syllable	Different first syllable
Common first syllable	**工作-工人** (gongzuo-gongren) *Work-Worker*	**老虎-时间** (laohu-shijian) *Tiger-Time*
Different first syllable	**名字-明天** (mingzi-mingtian) *Name-Tomorrow*	**厨房-熊猫** (chufang-xiongmao) *Kitchen-Panda*

Table 1. Example of Materials

Participants viewed two blocks of 100 word pairs presented in a randomized order. After a prestimulus interval of 200 ms, the first word was flashed for 500 ms at fixation followed by the second word after a variable interstimulus interval of 500, 600, or 700 ms. Participants were instructed to indicate whether the second word's sound was similar to the first one by pressing keys.

4. Conclusion

From the classic electrophysiological component N400, this experiment found that the English-Chinese bilinguals could not judge these Chinese word pairs easily when the result of explicit processing is not the same as the result of implicit processing. This finding provides that L1 can be activated and Stroop-like effects do exist in bilingual minds.

5. References

Hell, J. G. V., & Dijkstra, T. (2002). Foreign language knowledge can influence native language performance in exclusively native contexts. Psychonomic Bulletin & Review, 9(4), 780-789.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. Journal of Memory & Language, 33(2), 149-174.

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation during foreign-language comprehension. Proceedings of the National Academy of Sciences of the United States of America, 104(30), 12530-5.

Van Heuven, W. J. B., Schriefers, H., Dijkstra, T., & Hagoort, P. (2008). Language conflict in the bilingual brain. Cerebral Cortex, 18(11), 2706-2716.