RANLPStud 2015

Proceedings of the Student Research Workshop

associated with

The 10th International Conference on Recent Advances in Natural Language Processing (RANLP 2015)

7–9 September, 2015
Hissar, Bulgaria

STUDENT RESEARCH WORKSHOP
ASSOCIATED WITH THE INTERNATIONAL CONFERENCE
RECENT ADVANCES IN
NATURAL LANGUAGE PROCESSING'2015

PROCEEDINGS

Hissar, Bulgaria
7–9 September 2015

ISSN 1314-9156

Designed and Printed by INCOMA Ltd.
Shoumen, BULGARIA

Preface

The Recent Advances in Natural Language Processing (RANLP) conference, which is ranked among the most influential NLP conferences, has always been a meeting venue for scientists coming from all over the world. Since 2009, we decided to give arena to the younger and less experienced members of the NLP community to share their results with an international audience. For this reason, further to the first three successful and highly competitive Student Research Workshops associated with the conferences RANLP 2009, RANLP 2011, and RANLP 2013, we are pleased to announce the forth edition of the workshop which is held during the main RANLP 2015 conference days, 7–9 September 2015.

The aim of the workshop is to provide an excellent opportunity for students at all levels (Bachelor, Masters, and Ph.D.) to present their work in progress or completed projects to an international research audience and receive feedback from senior researchers. This year, we received 10 high quality submissions, among which 3 papers have been accepted for oral presentation, and 2 as posters. Each submission has been reviewed by at least 3 reviewers, who are experts in their field, in order to supply detailed and helpful comments. The papers' topics cover a broad selection of research areas, such as:

- application-orientated papers related to NLP;
- computer-aided language learning;
- dialogue systems;
- discourse;
- electronic dictionaries;
- evaluation;
- information extraction, event extraction, term extraction;
- information retrieval;
- knowledge acquisition;
- language resources, corpora, terminologies;
- lexicon;
- machine translation;
- morphology, syntax, parsing, POS tagging;
- multilingual NLP;
- NLP for biomedical texts;
- NLP for the Semantic web;
- ontologies;
- opinion mining;
- question answering;
- semantic role labelling;
- semantics;
- speech recognition;
- temporality processing;
- text categorisation;
- text generation;
- text simplification and readability estimation;
- text summarisation;
- textual entailment;
- theoretical papers related to NLP;
- word-sense disambiguation;

As usual, our authors comprise a large international group with students coming from: Bulgaria, Germany, India, and the United Kingdom.

We would like to thank the authors for submitting their articles to the Student Workshop, the members of the Programme Committee for their efforts to provide exhaustive reviews, and the mentors who agreed to have a deeper look at the students' work. We hope that all the participants will receive invaluable feedback about their research.

Irina Temnikova, Ivelina Nikolova and Alexander Popov
Organisers of the Student Workshop, held in conjunction with
The International Conference RANLP-15

Organizers:

Irina Temnikova (Qatar Computing Research Institute, HBKU, Qatar)
Ivelina Nikolova (Bulgarian Academy of Sciences, Bulgaria)
Alexander Popov (Bulgarian Academy of Sciences, Bulgaria)

Programme Committee:

Ahmed Abdelali (Qatar Computing Research Institute, HBKU, Qatar)
Corina Forascu (Alexandru Ioan Cuza University, Romania)
Goran Glavaš (University of Zagreb, Croatia)
Francisco Guzmán (Qatar Computing Research Institute, HBKU, Qatar)
Diana Inkpen (University of Ottawa, Canada)
Annie Louis (University of Edinburgh, United Kingdom)
Preslav Nakov (Qatar Computing Research Institute, HBKU, Qatar)
Constantin Orasan (University of Wolverhampton, UK)
Petya Osenova (Sofia University and IICT-BAS, Bulgaria)
Doaa Samy (Cairo University, Egypt)
Jan Šnajder (University of Zagreb, Croatia)
Stan Szpakowicz (University of Ottawa, Canada)
Andrea Varga (The Content Group Ltd, United Kingdom)

Table of Contents

The Complexity of Scrambling in Japanese: A TAG Approach

Alexander Diez
Heinrich-Heine University Düsseldorf, Germany
Alexander.Diez@hhu.de

Abstract

In this article, I present Japanese local and long-distance scrambling and restrictions to this phenomenon. I will argue that Japanese scrambling is too complex to be adequately represented with TAG. Instead, I will use a variant of TAG, namely TL-MCTAG. Subsequently, I also will propose to regard other scrambling languages, such as German, or Russian, in complexity classes, which is basically driven by the derivational power of each TAG formalism. This classification, though remains peripheric.

1 Introduction

This paper focuses on scrambling in Japanese, a language well known for its relative free word order. As a strict SOV language, Japanese verbs are linearized at the right end of a VP, while the other constituents may precede in any order without changing the denotation of the VP. This sort of flexiblity is known as scrambling (Bailyn, 2002, 83). There are roughly two types of scrambling, namely LOCAL SCRAMBLING (LS) and LONG DISTANCE SCRAMBLING (LDS). LS permits free word order inside the domain of a governing verb. Besides of the canonical order (1a), any permutation of the constituents is possible. (1b-c) shows some of the possible permutations.

(1) a. Hanako-ga hon-o
Hanako-NOM book-ACC
otōto-ni ageta
little.brother-to give.PST
'Hanako gave the book to the little brother.'

b. Otōto-ni Hanako-ga hon-o ageta.

c. Hon-o otōto-ni Hanako-ga ageta.

In contrast, LDS is highly restricted. LDS extends constituent boundary beyond the domain of a governing verb, as shown in (2b) in contrast to the canonical order in (2a).

(2) a. Kaito-ga [kinō Tarō-ga
Kaito-NOM yesterday Taro-NOM
Ginza-de sushi-to-sashimi-o
Ginza-in sushi-and-sashimi-ACC
tabeta] to Hanako-ni itta.
eat.PST COMP Hanako-to say.PST
'Kaito said, that yesterday Tarō ate sushi and sashimi in Ginza.'

b. Sushi-to-sashimi-o$_i$ Kaito-ga [kinō Tarō-ga Ginza-de t$_i$ tabeta] to Hanako-ni itta.

TREE ADJOINING GRAMMAR (TAG) (Joshi and Schabes, 1997) is a formalism based on tree-rewriting, where the so-called elementary trees can be combined by two compositional operations, substitution for initial trees and adjunction for auxiliary trees, and generate larger trees, as shown in Fig. 1. Both initial trees of *Peter* and *the fridge* rewrite the non-terminal leaves (substitution) of the *repaired*-tree. The auxiliary *easily*-tree rewrites the inner VP node (adjunction) of the *repaired*-tree. Auxiliary trees have a distinguished non-terminal leaf, the footnode, marked by an asterisk (*). After adjunction, the subtree of the rewritten target node appears below the footnode. The result of substitution and adjunction is a unique, derived tree. If every elementary tree includes at least one lexical anchor, the grammar is called lexicalized, hence a LEXICALIZED TAG (LTAG). It is, furthermore, possible to enrich the nodes of an elementary tree with feature structures, which is then called FEATURE STRUCTURE BASED TAG (FTAG) (Vijay-Shanker and Joshi, 1988). Such a unification based system is as powerful as TAG, but enhances the 'descriptive' capacity. Still, the generative power of

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 1–7,
Hissar, Bulgaria, 7-9 September 2015.

LTAG is not enough for Japanese LDS. This task, as will be shown in Section 3.3, can be solved with TREE-LOCAL MULTI-COMPONENT TAG (TL-MCTAG) (Weir, 1988). TL-MCTAG consists of sets of elementary trees, which must adjoin or substitute into one and the same elementary tree, and is equal to LTAG in terms of generative power. In section 3.3 I will show that TL-MCTAG, in contrast to LTAG, has the desired power to adequately analyse Japanese scrambling.

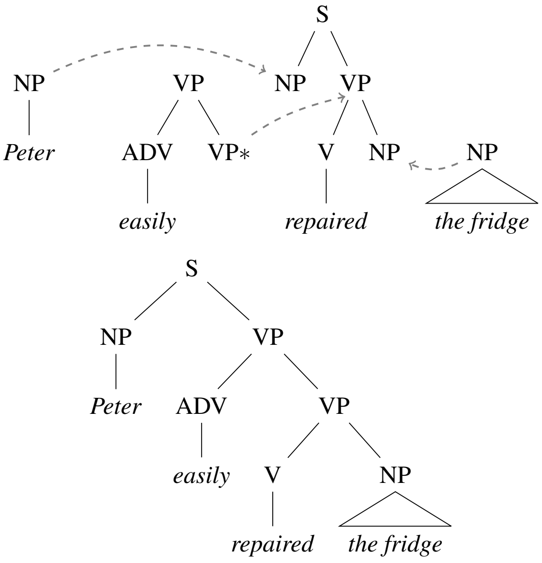


Figure 1: TAG derivation for the sentence *Peter easily repaired the fridge*.

One challenge lies in the complexity of Japanese scrambling, which can be shown with TAG. In this context, scrambling complexity refers to the generative power of the used TAG framework. In connection to this, the interesting question arrises how similar well-known scrambling languages such as German, Japanese, or Russian are and if they can be assigned to different complexity classes. I will argue that Japanese can be captured satisfyingly by TL-MCTAG, in contrast to German or presumably Russian. Becker et al. (1992) show that German LDS cannot be captured by TL-MCTAG and suggests a more powerful variant of TAG. Japanese allows LDS, as well, but it turns out to be more restricted. Russian, on the contrary, seems even less restrictive on scrambling compared to German (Sekerina, 2003).

In this work, I intend to provide LTAG analyses for scrambling in Japanese. As a result I will show, that LTAG is not powerful enough, whereas TL-MCTAG provides the desirable derivational power. Derivational power takes into account as to how derived structures are formed, instead of just regarding the derived structures themselves.

2 Word Order in Japanese

Tsujimura (2000) proposes six rules that restrict Japanese free word order. Firstly, as a strict head-final language, the verb cannot be scrambled and every constituent precedes the verb. (2a) shows the canonical order, while (3) opposes the first restriction.

(3) * Kaito-ga itta$_i$ [kinō Tarō-ga Ginza-de sushi-to-sashimi-o tabeta] to Hanako-ni t$_i$.

Secondly, the noun phrase and its corresponding particle are always considered a constituent and cannot be reordered separately (4). Also, particles always follow and assign the noun phrases, which is why constituents are represented as P(article) P(hrase)-leaves instead of NP-leaves. The examples (4)–(6) correspond to the glossing of (2a).

(4) * Kaito-ga [ga$_i$ kinō Tarō-$_i$ Ginza-de sushi-to-sashimi-o tabeta] to Hanako-ni itta.

Thirdly, the connective word *to* joins two or more NP to one constituent, so that it becomes ungrammatical to scramble only one of those conjoined NPs as in (5).

(5) * Kaito-ga [kinō Tarō-ga sashimi-o$_i$ Ginza-de t$_i$-sushi-to tabeta] to Hanako-ni itta.

Fourthly, scrambling out of the embedded clause, or rather out of the domain of the governing verb, also has its limits. Subjects and modifier are bound to their head and cannot be scrambled out of the verbal phrase. Otherwise, it leads to ill-formed sentences such as (6). Scrambling constituents other than the subject or a modifier, however, is possible and leads to LDS, see again (2b).

(6) * Ginza-de$_i$ Kaito-ga [kinō Tarō-ga t$_i$ sushi-to-sashimi-o tabeta] to Hanako-ni itta.

(2b) Sushi-to-sashimi-o$_i$ Kaito-ga [kinō Tarō-ga Ginza-de t$_i$ tabeta] to Hanako-ni itta.

Fifthly, and this is connected to the head-final structure, only leftward movement is allowed, and thus, extraposition is ruled out. Finally, particle order may forbid scrambling, as well. For instance, a *ga-ni* particle order is free for scrambling, while a *ga-ga* order forbids LS of both constituents. The particle *ga* usually is a subject marker.

For LDS, two more conditions are required: complementizing and a governing verb of perception. In a variety of articles (Saito, 2012; Suzuki, 1994, among others) examining Japanese free word order, it is stated that the *to*-complementizer is prominent to permit scrambling out of sentence boundaries. The *to*-complementizer marks direct and indirect speech, and thus will be denoted as a QUOTATION PARTICLE (PQ) POS-TAG. In fact, there exist more complementizers which permit LDS, i.e., the FORMAL NOUNS (NF[1]) *koto*, and *no*, or *yō* and *ka*. Together with these complementizers, verbs like *iu* ('say'), *omou* ('think'), *shiru* ('know'), or *kangaeru* ('reason') allow for LDS.

Furthermore, complementizing can be divided into two groups of precedence patterns. In both structures the complementizing (COMP) part is preceded directly by the verb of the embedded clause (V2). Then, the governing verb (V1) succeeds the complementizer, but also allows constituents, a dative case for instance, in between. Yet, *to* or *ka* complementizer do not demand any particles (P), as suggested in (7). A NF, however, requires an immediately succeeding particle, since it nominalizes the relative clause. This precedence structure is shown in (8). Strict precedence is marked by $>$, whereas $\gg$ is non-strict precedence.

(7) V2 $>$ COMP $\gg$ V1

(8) V2 $>$ NF $>$ P $\gg$ V1

(7) and (8) are crucial for the TAG analysis proposed in Section 3.2, insofar that V1, V2, COMP, NF, and P will appear as leaves in the elementary trees.

3 TAG Modelling

3.1 Underlying Linguistic Principles

TAG is a mathematical formalism in the first place. It lacks the linguistic interpretation, e.g., principled constraints on the shape of elementary trees of the nodes and syntactic structure. In this article, I use the valency principle according to Frank (2002) and Lichte (to appear, Section 5.3) as linguistic interpretation. In short, the lexical anchor represents the valency head or carrier and the non-terminal leaf nodes the valency roles. Modification, on the other hand, is factored away into a separate elementary tree (such as *easily* in Fig. 1). Still, as Frank (2002, 22) points out as the Fundamental TAG Hypothesis, every syntactic dependency, such as valency relation, is expressed locally within an elementary tree. The valency head co-occures with its arguments, which are non-terminal leaves (such as both NP-nodes of *repaired* in Fig. 1). Functional trees do not realize any valency and thus, the valency principle does not apply (Lichte, to appear, Section 5.3).

3.2 Scrambling with LTAG

Each instance of LS can be easily linearized in one elementary tree, as can be seen from the elementary trees in Fig. 2. The valency carrier (head) *ageta* ('gave') has three valency roles realized in one elementary tree. For each linearization the PP leaf nodes take constituents with the fitting particles. The constituents, which substitute into the PP leaf nodes, can be realized at only that position.

Realizing LDS with LTAG needs a more elaborate approach, since scrambling outside the domain of a head needs a particular sentence structure. This condition is realized with the PQ and NF elementary trees in Fig. 3, and in accordance to the precedence relations proposed in (7) and (8). The first requirement to these auxiliary trees is to permit embedding of further sentences. Verbs like *omou* ('think') would need a NF node, which enables embedding, while verbs like *iu* ('say') need a quotation particle *to*. Furthermore, a NF itself needs a succeeding particle, while PQ does not, resulting in slightly different auxiliary trees, as shown in Fig. 3. The initial tree and auxiliary tree of *itta* ('said') are used for the derivation in Fig. 4.

Note that the number of PP-leaves may differ as they represent valency roles. The VP-footnode adjoins to the initial tree, which is supposed to be the embedded tree after derivation. In addition, and on the basis of FTAG, VP nodes of the elementary trees carry feature structures with a Boolean SUBJ $+$ attribute, in contrast to the VP-foot node, which is enriched with SUBJ $-$. The PP that hosts a subject passes this information on to the VP-child of his VP-sister. This mechanism makes sure that no auxiliary tree with a

[1] This term was introduced in the Japanese treebank of the VERBMOBIL project (Kawata and Bartels, 2000, 28). The semantic content of formal nouns is empty and they are used to form nominal structures together with other expressions.

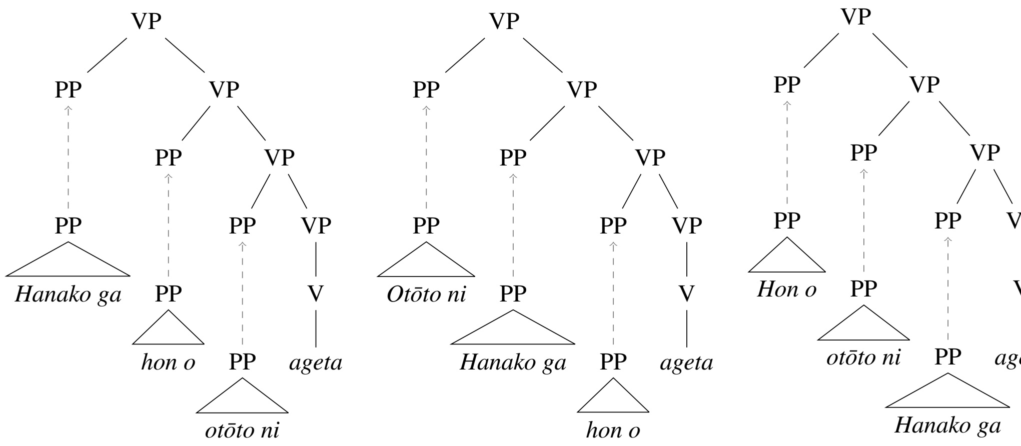
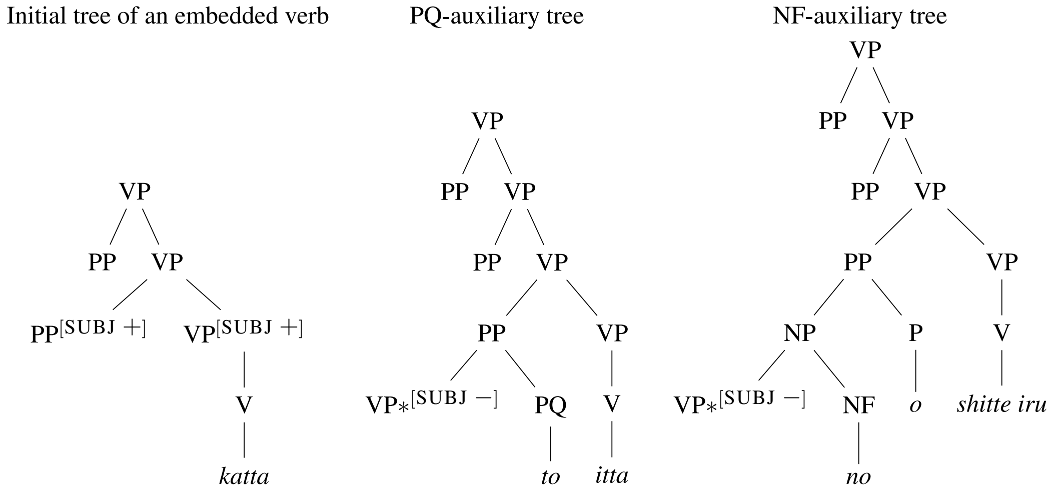


Figure 2: Elementary trees for the local scrambling of arguments of *ageta* ('give'). They are needed to derive the sentences in (1).


Initial tree of an embedded verb

PQ-auxiliary tree

NF-auxiliary tree

Figure 3: Elementary trees: initial tree, PQ- and NF-auxuliary trees.

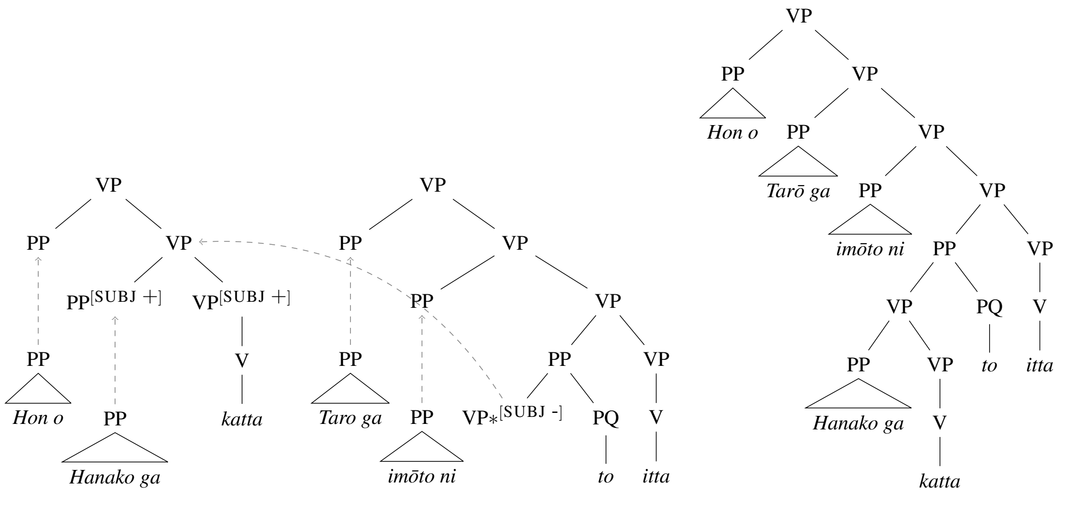


Figure 4: TAG derivation of (9), where *hon o* ('book', accusative) undergoes LDS.

SUBJ −-marked footnode can adjoin to a (partial) tree, which is dominated by a leaf with a SUBJ +-marking, and thus, restrict long-distance scrambling. An LTAG derivation of sentence (9), where *hon o* ('book', accusative) is the long distance scrambled constituent, is shown in Fig. 4. The sentence is glossed with the corresponding string schema (Becker et al., 1991), where $N1$ and $N2$ are valency roles of $V1$ and $V2$, respectively.

(9) Hon-o Tarō-ga imōto-ni Hanako-ga katta
N2 N1 N1 N2 V2
to itta.
V1
'Taro told to the little sister that it is the book, which Hanako bought.'

However, LTAG is not sufficient to fully cover LDS in Japanese. The auxiliary tree of Fig. 4, which has the string schema $N1$ $N1$ $V1$, where N corresponds to the valency role of the head V, adjoins to the initial tree ($N2$ $N2$ $V2$), resulting in the string schema $N2$ $N1$ $N1$ $N2$ $V2$ $V1$.[2] The derivational power (Becker et al., 1991) of LTAG meets its limits with the sentence structure $N1$ $N2$ $N1$ $N2$ $V2$ $V1$. From the derivational point of view, there is no possibility to generate such a structure with tree templates as in Fig. 3. Prospective new trees (see Fig. 5)[3] would need additional inner nodes for adjunction and also the relation between auxiliary tree and initial tree would be changed, since the footnode is in the primary initial tree. Thus, the tree modelling would be inconsistent, and additional inner nodes in the initial tree for the sake of adjunction would become necessary. On the other hand, auxiliary trees would lack non-terminal leaves. Hence, the valency principle would be violated. Also, the projection of the complementizer, which has to be realized as an auxiliary tree, contradicts the valency principle by being realized in the initial tree.

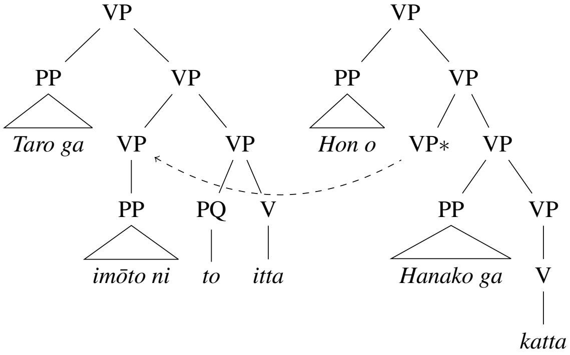


Figure 5: Contradiction of the Valency Principles: new trees for deriving the string schema $N1$ $N2$ $N1$ $N2$ $V2$ $V1$.

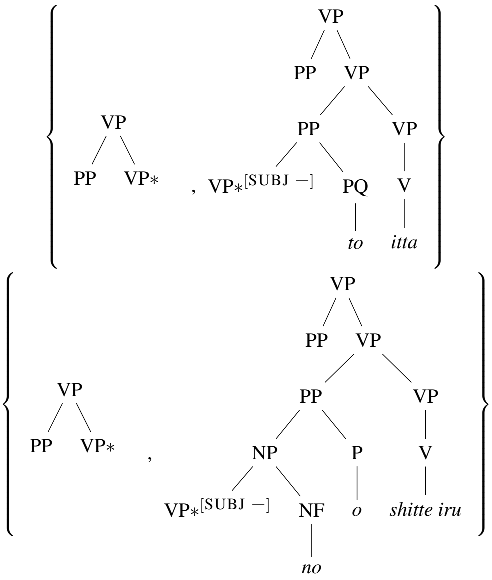


Figure 6: MC-Tags of PQ- and NF-trees from Fig. 3.

3.3 TL-MCTAG: Gaining more Derivational Power

Chen-Main and Joshi (2014) propose TREE-LOCAL MULTI-COMPONENT TAG (TL-MCTAG), which is equal to TAG in generative power, but more powerful in derivational terms, for ill-nested dependency structures or gap degree > 1. TL-MCTAG, as proposed in Fig. 6, permits to integrate the auxiliary trees of Fig. 3 as MC-sets. Doing so, string schemas such as $N1$ $N2$ $N1$ $N2$ $V2$ $V1$ can be generated without contradicting the valency principles other than the TAG analysis in Fig. 5. Fig. 7 proposes the derivation of the string schema in Fig. 5 with TL-MCTAG.

Furthermore, string schemas, which could be problematic for TL-MCTAG, do not appear in Japanese. For instance, since extraposition is ruled out, there is no possibility of a valency role preceding the head. Also, even if constituents are scrambled according to the restrictions in Section 2, LDS of more than one constituent results in un-

[2] Note that the sentence structure $N2$ $N1$ $N2$ $V2$ $N1$ $V1$ also lies within the derivational power of LTAG.

[3] The meaning of this sentence is slightly different: 'Taro told the little sister, that it is a book which Hanako bought'

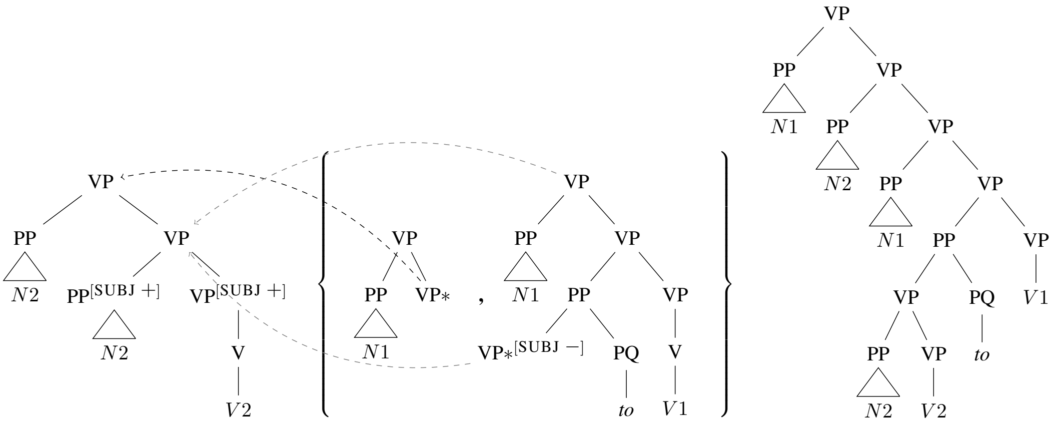


Figure 7: TL-MCTAG derivation of the string schema $N1$ $N2$ $N1$ $N2$ $V2$ $V1$.

grammatical sentences. The word order in (10a), with the scrambled *Hisarya-e*, is a grammatical sentence in Japanese, while the additional scrambling of *hon-o* in (10b) results in an ungrammatical sentence.[4]

(10) a. Hisarya-e Tarō-ga Ichirō-ni
Hissar-to Tarō-NOM Ichirō-DAT
Hanako-ga hon-o okutta
Hanako-NOM book-ACC send.PST
to itta
COMP say.PST
'Taro told Ichiro, that it is Hissar, where Hanako send the book.'

b. * Hisarya-e Tarō-ga hon-o Ichirō-ni
N2 N1 N2 N1
Hanako-ga okutta to itta
N2 V2 V1

4 Complexity Comparison

With TAG it is possible to compare language complexity, considering scrambling complexity, in particular. As discussed in this paper, a fully sufficient TAG variant for Japanese, in the terms of generative and derivational power, is TL-MCTAG. Becker et al. (1992) show, however, that TL-MCTAG does not suffice for German, since scrambling (LS and LDS) is unbound to distance and number of dependencies. Russian, on the other hand, can only be vaguely categorized, due to a lack of literature that discuss Russian LDS exhaustively. Glushan (2006) shows that Russian LDS has similarities with Japanese LDS, but is less restricted. For instance, in Russian it is permitted in some cases to dislocate subjects and modifier out of embedded clauses. Yet, Glushan (2006) points out that scrambling is also more restricted than assumed in the literature, since she could list a number of cases, where LDS is either unrestricted or clearly restricted. Additionally, she argues that Russian scrambling can be successive cyclic. It is unclear, though, whether 'doubly unbounded' constructions are possible, similar to German. Thus, Japanese scrambling is less complex than German, since TL-MCTAG is sufficient for Japanese but not for German anymore. Russian appears to be more complex than Japanese, but it remains unclear if the complexity is lower, equal, or even higher than German.

5 Conclusion

In this paper I have shown that scrambling in Japanese, even though being very flexible within local domains, underlies considerable constraints when it becomes non-local. Eventually, these constraints require some derivational power (in terms of string schemata) that is already available in TL-MCTAG. This result is in sharp contrast to TAG-approaches to scrambling in other languages, notably German (Becker et al., 1992), where much more powerful extensions of TAG are necessary. Another language of this sort seems to be Russian (Sekerina, 2003). It therefore seems that scrambling cross-linguistically falls into different complexity classes, which can be neatly characterized within the TAG-framework.

References

John Frederick Bailyn. 2002. Scrambling to reduce scrambling. *Glot International*, 6(4):83–90.

[4] I'm grateful to Mamoru Saito for helping me with this sort of sentences

Tilman Becker, Aravind K. Joshi, and Owen Rambow. 1991. Long-distance scrambling and Tree Adjoining Grammars. In Jürgen Kunze and Dorothee Reimann, editors, *EACL '91 Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, pages 21–26.

Tilman Becker, Owen Rambow, and Michael Niv. 1992. The Derivationel Generative Power of Formal Systems or Scrambling is beyond LCFRS. Technical Report IRCS-92-38, Institute for Research in Cognitive Science, University of Pennsylvania.

Joan Chen-Main and Aravind K. Joshi. 2014. A dependency perspective on the adequacy of Tree Local Multi-Component Tree Adjoining Grammar. *J. Log. Comput.*, 24(5):989–1022.

Robert Frank. 2002. *Phrase structure composition and syntactic dependencies*, volume 38 of *Current studies in linguistics*. MIT Press, Cambridge and Mass.

Zhana Glushan. 2006. *Japanese Style Scrambling Russian: Myth and Reality*. Ph.D. thesis, University of Tromsoe.

Aravind K. Joshi and Yves Schabes. 1997. Tree-Adjoining Grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 3, pages 69–124. Springer, Berlin, New York.

Yasuhiro Kawata and Julia Bartels. 2000. Stylebook for the Japanese treebank in Verbmobil.

Timm Lichte. to appear. *Syntax und Valenz: Zur Modellierung kohärenter und elliptischer Strukturen mit Baumadjunktionsgrammatiken*. Language Science Press, Berlin.

Mamoru Saito. 2012. Sentence types and the Japanese right periphery. In G. Grewendorf and T. E. Zimmermann, editors, *Discourse and grammar*, volume 112 of *Studies in generative grammar*, pages 147–176. de Gruyter Mouton, Boston and Berlin.

Irina Sekerina. 2003. Scrambling and processing: Dependencies, complexity, and constraints. In Simin Karimi, editor, *Word order and scrambling*, volume 4 of *Explaining linguistics*, pages 301–324. Blackwell Pub., Malden and Mass.

Satoko Suzuki. 1994. That a fact? Reevaluation of the relationship between factivity and complementizer choice in Japanese: Proceedings of the twentieth annual meeting of the berkeley linguistics society: General session dedicated to the contributions of charles j. fillmore. *The Berkeley Linguistics Society*, 20(1):521–531.

Natsuko Tsujimura. 2000. *An introduction to Japanese linguistics*, volume 10 of *Blackwell textbooks in linguistics*. Blackwell, Cambridge and Mass, repr. edition.

K. Vijay-Shanker and Aravind K. Joshi. 1988. Feature structure based Tree Adjoining Grammars. In *In proceedings of COLING*, pages 714–719.

David J. Weir. 1988. *Characterizing mildly context-sensitive grammar formalisms*. Ph.D. thesis, University of Pennsylvania.

Some Theoretical Considerations in Off-the-Shelf Text Analysis Software

Emma Franklin
University of Wolverhampton
emma.franklin@wlv.ac.uk

Abstract

This paper is concerned with theoretical considerations of commercial content analysis software, namely Linguistic Inquiry and Word Count (LIWC), developed by social psychologists at the University of Texas. LIWC is widely cited and forms the basis of many research papers from a range of disciplines. Here, LIWC is taken as an example of a context-independent, word-counting approach to text analysis, and the strengths and potential pitfalls of such a methodology are discussed. It is shown that text analysis software is constrained not only by its functions, but also by its underlying theoretical assumptions. The paper offers recommendations for good practice in software commercialisation and application, stressing the importance of transparency and acknowledgement of biases.

1 Introduction

Due to the ever-increasing availability of digital texts, automated text analysis methods are gaining popularity, and not only among linguists. Commercial text analysis programs can offer fast, inexpensive content analysis and are generally quite easy to use. To a social scientist faced with large amounts of discourse to analyse, such a product is almost irresistible. However, there are several caveats of which end-users should be made aware, and to which software developers could give more explicit attention.

There are a number of user-friendly and easily obtainable text analysis programs currently on the market. One which has been described as the "most widely used program for analysing text in clinical psychology" (Alpers et al., 2005: 363), and which continues to grow in popularity across a number of fields, is Linguistic Inquiry and Word Count[1] (LIWC, pronounced 'Luke'). This paper takes LIWC as an example of commercially available, easy-to-use text analysis software. It aims to illustrate the ways in which text analysis software is tightly bound to its theoretical basis, and the practical implications this can have on usage and performance. We find this a timely discussion due to the diverse range of researchers now turning to linguistic analysis software without necessarily understanding its construction.

The paper is organised as follows. Section 2 outlines the LIWC system and its development, followed by some of its perceived theoretical assumptions in Section 3. Section 4 describes some previous experiments using LIWC in Franklin (2015). Section 5 revisits LIWC's theoretical assumptions, and offers suggestions for good practice in software commercialisation and application. Section 6 concludes the paper.

2 LIWC

First developed by social psychologists at the University of Texas in the 1990s, and now in its second version, LIWC2007 is described as "a transparent text analysis program that counts words in psychologically meaningful categories" (Tausczik and Pennebaker, 2010: 24). It was originally employed in social psychology for investigating the connections between word use and mental health recovery (Pennebaker, 1997), later followed by relationship satisfaction (Agnew et al., 1998), university grades (Pennebaker and King, 1999), and testosterone levels (Pennebaker et al., 2004), amongst others. It has also been used to track collective responses to upheaval, such as the 9/11 attacks (Cohn et al., 2004).

The creators of LIWC have since used the program to analyse the essays, speeches, blog posts and Tweets of thousands of individuals across a variety of projects (Pennebaker, 2011). This has led to many others carrying out LIWC-based research in a range of applications: personality profiling (Mairesse and Walker, 2006), deception detection (Pérez-Rosas et al., 2014), sentiment analysis (Paltoglou et al., 2010), content analysis (Tumasjan et al., 2010) and review spam detection (Ott et al., 2013), to name a few.

[1] http://www.liwc.net

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 8–15,
Hissar, Bulgaria, 7-9 September 2015.

2.1 Functions, Development and Validation

LIWC performs a simple word-count analysis by reading text files and matching each word against its inbuilt categories, or dictionaries. These categories – of which there are 68 – each constitute a 'dimension' of words, e.g. positive emotion words, prepositions, motion words, and so on.[2]

Some dimensions attempt to describe themes, or content (e.g. 'work'), while others count grammatical features (e.g. 'verbs'). For each text file, LIWC generates a score for each dimension, which reflects the percentage of the total words which match that category. So, a score of 3.9 for 'past' would indicate that 3.9% of that text consists of words which can be found in the LIWC 'past' dictionary. This allows the user to easily track changes in LIWC scores over time. A word can belong to more than one category; 'died', for example, appears in 'past', 'verbs' and 'death'.

The dictionaries are "the heart of the LIWC program" (Tausczik and Pennebaker, 2010: 27), and were finalised over the course of several years. The dictionaries were populated with the help of existing resources (e.g. Roget's Thesaurus) and "brain-storming sessions among 3-6 judges" (Pennebaker et al., 2007: 7). The word lists were then rated by three judges, who voted on whether each word should be used in that category, and whether new words should be added; two out of the three judges needed to be in agreement for a decision to be passed. The refined dictionaries were then rated once more by three different judges, with the same criteria for selection and deletion. Inter-judge agreement "ranged from 93% agreement for Insight to 100% agreement for Ingestion, Death, Religion, Friends, Relatives, and Humans" (Pennebaker et al., 2007: 7). LIWC was appraised again in 1997 and 2007 (Pennebaker et al., 2007).

External validation of the dictionaries was carried out in Pennebaker and Francis (1996) for a select number of categories, by asking four judges to rate student essays against LIWC-compatible dimensions. The judges' ratings were reasonably correlated with the corresponding LIWC categories, although different correlation scores are reported in Pennebaker et al. (2007), presumably due to the revision of the LIWC dictionaries. The (Pearson) correlation coefficient for these few categories vary from 0.07 for Sadness to 0.87 for Family. The authors conclude that these results give "support for LIWC's external validity" (Pennebaker et al. 2007: 9).

LIWC has been generally well received, and its lexicon has been described as "the standard for social psychological analysis of lexical features" (Jurafsky et al. 2009). Some scholars explicitly note the effectiveness of LIWC's simplistic approach (Mehl, 2006), while others find that LIWC performs well, but only in certain categories and for certain domains (Loughran and McDonald, 2011; Pérez-Rosas et al., 2014; Grimmer and Stewart, 2013). In some studies, only the dictionaries from LIWC are adopted (Skowron and Paltoglou 2011, Bae and Lee, 2012), and in others the authors create their own dictionaries to be used with the LIWC processor (Loughran and McDonald, 2011; Osherenko and André, 2007). To a social scientist looking at language, LIWC might seem an obvious choice of analysis tool.

3 Theoretical Assumptions

Text analysis programs are inherently philosophical. That is to say, all software is underpinned by theories and assumptions, and in the case of text analysis software, these are theories of language and assumptions on what constitutes meaning. This may seem obvious, but the implication is that the end user of a program also, perhaps unknowingly, applies these theories to their research, unless they are able to fully understand the way the program works and account for these effects later. Therefore, it is crucial that we identify and critically assess the theoretical underpinnings of a text analysis program before using it.

Below are some of LIWC's main features and its perceived theoretical assumptions (numbered in brackets).

- **LIWC counts words.** Based on the context in which LIWC was created, the underlying assumption (A1) is that the frequency of a word can tell us something about a person or about the content or tone of a text. A secondary assumption (A2) is that a computer program is ideal for carrying out this task.
- **It only considers single words.** In doing so, LIWC assumes that words have meaning in isolation (A3). There is also an implicit assumption (A4) that inaccuracies due to negation, word order, particles (e.g. in phrasal verbs), ambiguity of word senses, type of discourse and other context-dependent factors are negligible or unimportant.

[2] The full list of LIWC categories is available at http://www.liwc.net/descriptiontable1.php

- **It matches words against dictionaries.** The assumption here (A5) is that by creating dictionaries *a priori*, and by finding and counting only the words which match them, the interesting or relevant parts of a text will be identified. A secondary assumption (A6) is that a score corresponding to a dictionary label will also correspond to the intended semantic value of that dictionary. Thirdly, it is assumed that dictionary values are more useful than word frequencies (A7).

Of course, there are many text analysis approaches which share several of these assumptions, and LIWC is not an unusual case, although it is particularly insensitive to context.

Section 4 describes some experiments carried out by Franklin (2015) (based loosely on Cohn et al.'s (2004) LIWC analysis of American blogs written before and after the 9/11 attacks). Results from the study, which examined the transition experienced by new university students and the different outcomes of LIWC and keywords analyses, are selected so as to address the assumptions listed above (A1-7) as succinctly as possible.

4 Word-Count Software in Practice

Franklin (2015) sought to better understand the changes undergone by first-year university students following the move to university, with particular focus on student identity and preoccupation. The study was also an investigation into the efficacy of a word-count approach compared with more manual corpus analysis methods. Taking as data the blog posts of thirty new students in the two months preceding and following the move to university, language changes over this period were examined. A LIWC analysis was carried out, using all of the standard LIWC2007 dictionaries, followed by a log-likelihood keywords analysis. In both cases, Corpus B (blogs written after the move) was compared against Corpus A (blogs written before the move). Results were examined manually using concordancer AntConc (Anthony, 2011). Corpus details are given below.

Corpus	**Tokens** (types)	**Total**
A: Blogs written before moving	232,242 (14,248)	389,721
B: Blogs written after moving	157,479 (10,536)	

Table 1: Corpus details

LIWC scores (for all 68 dimensions) were generated for each student's 'before' and 'after' blog posts. 'Change scores' were then calculated for each student, in each LIWC dimension, by dividing the LIWC scores for all of their entries written after the move to university by the LIWC scores for their entries written beforehand, then subtracting 1. This produced a negative score (a drop in LIWC score), a neutral score (no change), or a positive score (an increase in LIWC score). Overall LIWC change scores were then calculated for each category by subtracting the number of people for whom the change was negative from the number of people for whom it was positive. This was carried out three times, with different thresholds[3], and then the scores averaged. This final score was used to rank the LIWC dimensions and determine the categories, or dimensions, whose scores changed the most overall.

A keywords analysis was then carried out on Corpus B, using Corpus A as the reference corpus. Finally, the results for the LIWC analysis and keywords analysis were compared.

4.1 Findings

Table 2, below, gives the fourteen LIWC categories with the greatest overall change across all students, be it positive (+) or negative (-). However, the problem with results such as these is that they do not illustrate actual changes in word use. For function-word categories such as 'we', whose dictionaries contain a small number of unambiguous words, the LIWC score can paint a reasonably clear picture of general language changes. For larger, vaguer categories such as 'leisure', 'health' and 'religion', however, the scores alone cannot realistically convey what is happening in the data.

Categories with the greatest change scores			
future	-16.00	***filler***	+11.33
we	+15.67	***humans***	+10.33
see	-15.33	***health***	+9.33
leisure	-14.33	***excl***	+9.00
assent	+13.67	***cogmech***	+9.00
number	+12.67	***relig***	+9.00
motion	-11.33	***preps***	-8.67

Table 2: Categories with greatest change scores

[3] First, taking 0 as the threshold, i.e. anything above 0 was considered a positive change and anything below 0 a negative change; then with a threshold of -/+0.5; then with a threshold of -/+1. This was done to account for both the strength and the breadth of the LIWC changes.

The AntConc concordancer was used to examine the LIWC words in context, which helped to explain the results. The drops in ‘future’ and ‘motion’ scores, for example, were corroborated by the concordance lines; before university, the students were anticipating the move and used words such as ‘gonna’, ‘will’, and ‘leave’, which decreased once the move date had passed. Increases in ‘number’ and ‘humans’ scores were also predictable; the students are now *first*-year students, meeting *people* and joining *societies*. LIWC was also correct to identify a greater ‘health’ preoccupation; the new students were reportedly tired, hung-over and suffering from ‘freshers’ flu’.

The ‘see’ category score, however, was highly skewed by mentions of the word ‘looking’, as used in ‘looking forward’ [to university], which dropped following the move. This was therefore a somewhat misleading score change, since the students did not appear to be ‘seeing’ less − at least, not in the literal sense. However, a concordance analysis revealed some interesting changes in *how* they saw things; the construction *LOOK + adj* tended to feature quite general, positive adjectives before the move (e.g. ‘good’, ‘nice’), with slightly more specific, critical adjectives being used after the move (e.g. ‘weird’, ‘edgy’).

The drop in the ‘leisure’ score suggested that the students were now engaging in fewer leisure activities, which may have been true, given their busy university schedules. However, this drop in score was also masking some *increases* in leisure words. The word ‘reading’, for example, was found to be used more frequently after the move. Going on word frequency alone, this might lead the researcher to assume that academic reading had become a greater preoccupation. However, on examining the context, a main cause was found to be the students’ mentions of ‘reading’ with relation to their own blogs, which increased by almost half. A concordance analysis found that the students became increasingly concerned with the impressions they gave to readers, something which could not be identified using LIWC alone.

Increases in ‘assent’ and ‘filler’ were interesting, as these categories were meant for transcripts of spoken language. The results were characterised by word sense errors, namely the adjective ‘cool’ in the ‘assent’ category, and the verb ‘like’ in ‘filler’, but investigations into these categories using the concordancer still yielded useful findings: students were using words such as “yeah” and “so yeah” to relate to the reader, and “feel like” and “it’s like” to describe their new university experiences. From this, and other findings, it was discovered that the bloggers displayed a greater preoccupation with their readership after the move to university. In this case, LIWC played a pivotal role in prompting this line of inquiry.

The most effective LIWC category was ‘we’, which made it possible to reliably track all mentions of first-person plural pronouns (though the referents of the pronouns had to be manually identified). Despite not being able to tell us to whom these pronouns referred, this small, closed-class category proved useful in measuring a sense of inclusiveness and collective identity. The fact that this dictionary is unlikely to be affected by noise and ambiguity made it possible to plot each student’s individual ‘we’ scores on line graphs, demonstrating the rises and falls in these ‘we’ words on a post-by-post basis, over time.

The increase in the ‘religion’ score was of particular interest in the context of this study, as the literature suggests that students who move away for university tend to become *less* religious (Bryant et al., 2003). On closer examination it was found that the increase was mostly due to noisy matches such as ‘seminar’ (due to the inclusion of *seminar**, intended to match ‘seminary’ and ‘seminaries’). Further erroneous matches were ‘demonstration’ (from *demon**), ‘scuba diving’ (*divin**) and ‘monkeys’ (*monk**). There were also a number of ‘religious’ words which were actually not religious in the context of student blogs (e.g. ‘Christmas’ as an end-of-term marker as opposed to religious holiday). In fact, when all LIWC 'religion' hits were manually checked, it was found that there was not an *increase* in religious uses of these terms, but a *decrease*.

When compared against the findings yielded by a keywords analysis, there was high overlap; out of the 38 findings of the study, 25 were shared by both the LIWC and keywords analyses. However, significantly more time was spent on 'unravelling' the LIWC results than those generated by the keywords, as some of the LIWC words triggered misleading categories due to contextual or morphological inaccuracies. For both LIWC and keywords, however, a manual examination of the context was crucial; out of all 38 findings, 28 relied upon the consideration of context. See Table 3 in the Appendix for a list of all findings.

5 Discussion

5.1 Theoretical Assumptions Revisited

Taking some of the above findings as examples, and drawing on other examples where relevant, the validity and implications of assumptions A1-7 from Section 3 are now discussed.

A1: *the frequency of a word can tell us something about a person or about the content or tone of a text.*

Several psychological studies have used word frequencies to show correlations between word use and the mind, due to latent, albeit crude, associations with words (Rosenberg, 1990; Mehl, 2006). The bag-of-words approach has been taken by many researchers in other fields, too; Biber (1988), for example, has successfully used word frequencies to discriminate text type and genre. Word frequencies were certainly useful in the student study, but had to be examined in context.

A2: *a computer program is ideal for counting words.*

Computers are undoubtedly more efficient at counting than humans. In the context of psychoanalysis, it has also been argued that computers are better at seeing 'past' meaning and counting the less interesting but nonetheless relevant language patterns to which a human annotator might be desensitised (Spence, 1980).

A3: *words have meaning in isolation.*

This assumption is problematic – or, in the view of Hanks (2013), false. Words, he argues, do not have meaning, but *meaning potential*; their meanings can only be activated by context. This is not to say that single words cannot act as discriminating features of texts, but that semantic value cannot legitimately be ascribed to them.

Words which are less affected by this problem are closed-class, i.e. function words. This would explain why, out of all of the categories analysed in Section 4, the 'we' category was found to be the most accurate and reliable. It might also explain why there are many successful LIWC studies concerning pronoun use (Pennebaker, 2011).

A4: *inaccuracies due to context-dependent factors are negligible or unimportant.*

The justification for a context-independent system is that a word-count program is probabilistic, and therefore such inaccuracies are, statistically, so rare that they do not impact on results in a serious way. This is probably true, overall, when considering all LIWC features together, due to high accuracy rates in some categories. However, there are some categories and domains for which this effect is particularly strong and *does* affect the results in a serious way. Bond and Lee (2005), for example, found LIWC to be reasonably accurate, but not accurate enough to be used in "high-stakes" investigations; in their study of deceptive statements, 30% were classified incorrectly.

It has also been argued that a general-purpose dictionary such as LIWC's cannot be accurately applied to all domains and discourses. Loughran and McDonald (2011), for example, found that when using the Harvard IV dictionary (a lexicon similar to that of LIWC), three quarters of all words classified as 'negative' were in fact not negative in the *context* of the financial domain, just as many 'religious' words were not religious in the *context* of student blogs. Again, such levels of inaccuracy could not be considered negligible.

A5: *by creating dictionaries* a priori, *and by finding and counting only the words which match them, the interesting or relevant parts of a text are identified.*

It is worth mentioning that 'religion', the category which suffered the most from inaccuracies in the student study, was one of the few dictionaries reported as having "100%" inter-judge agreement. We know therefore that 100% inter-judge agreement (between two judges) does not guarantee a well-compiled dictionary. But even if a content-word dictionary were impeccably constructed, with high agreement among hundreds of judges, it would still have the problem of being subjective and culture-specific (Mehl, 2006). A dictionary-based approach to text analysis therefore suffers from two biases: first, the top-down, pre-defined nature of its word-matching process (as opposed to a bottom-up, data-driven, inductive approach); and secondly the bias that comes with domain-specific, culture-bound dictionaries.

A6: *a score which corresponds to a dictionary also corresponds to the intended semantic value of that dictionary.*

Due to context- and dictionary-related problems, some categories used in the student study provided misleading scores, reflecting instead an increase in the use of words which were indicative of some other topics or events. A cursory glance at the LIWC scores, without actually looking at the text (which is what many LIWC analyses consist of), might lead a researcher to falsely conclude that moving to university is associated with becoming more religious or seeing less, for example.

A7: *dictionary values are more useful than single word frequencies.*

A problem encountered with LIWC, and presumably other dictionary-based approaches, is that dictionary scores do not tell us the actual linguistic changes that have occurred. Instead, we are given a simple numerical output. Despite being described as 'transparent', LIWC is, in this sense, surprisingly opaque.

The main issue with this approach, however, is that a LIWC score can theoretically mean nothing. Two texts might have the same LIWC score in the same dimension, and yet be made up of completely different words. Secondly, as in the case of 'leisure' in Section 4, a LIWC category change score might show an overall change in one direction, while simultaneously masking the opposite change for particular words within that category. Such problems are, fortunately, easy to overcome with the use of complementary qualitative analysis tools, such as a concordancer.

On the other hand, there are dictionary values which are arguably more useful than individual word counts. The 'we' category, for example, made it possible to track mentions of collective identity over time, something which would have been far less convenient to do otherwise.

5.2 Recommendations for the Use and Development of Text Analysis Software

There are two parties involved in any software use: the developer, and the end user. We therefore propose two main courses of action in order to maximise the benefits and avoid the pitfalls of an off-the-shelf text analysis package such as LIWC.

The developer could:

1. Formulate a list of the main analytical functions of the program and their perceived theoretical assumptions, as is done in Section 3 of this paper. Our own assumptions can be hard to determine without the help of others, so this should be a collaborative, peer-reviewed effort. This will help the developer to identify any potentially problematic assumptions embedded in their software.
2. Publicise the above information as a clear and concise "readme" document, along with the usual user manual and validation papers. This would ensure that the end user, whose background may be in an unrelated discipline, is easily made aware of the potential philosophical biases and constraints of the software, rather than simply knowing how to install and run it.
3. Attempt to avoid dictionary-related problems by thoroughly checking their contents for morphological errors and likely ambiguity. Employ raters from a range of cultural and educational backgrounds and ensure that at least one linguist is involved in the creation and validation of such modules.
4. Try to use bottom-up, data-driven approaches to dictionary population, if applicable.

The end user could:

1. First assess their own research needs and their existing theoretical assumptions, and to make sure that the software they choose is in line with those. Of course, this is only possible if the program's theoretical and philosophical underpinnings have already been established.
2. Combine top-down, pre-defined, quantitative analytical approaches with more bottom-up, inductive, qualitative approaches. This will add depth to findings and avoid misleading dictionary scores being taken at face value.
3. Favour smaller dictionaries with closed-class words, i.e. pronouns and function words, which tend to be less ambiguous in meaning.
4. Prioritise context: if necessary, create a customised, domain-specific dictionary suited to the research area; and always examine results in context, e.g. by using a concordancer.

6 Conclusions

This paper used the program Linguistic Inquiry and Word Count (LIWC) to exemplify some of the main advantages and potential pitfalls of off-the-shelf text analysis software. Given the growing popularity of computerised text analysis, it is important that reductive, word-count programs such as LIWC are used with caution, particularly by researchers outside of linguistics and natural language processing. It should be made especially clear to users that, far from being 'objective' and philosophically neutral, all computer software is based on theoretical assumptions, some of which are the subjects of ongoing debate.

As regards the efficacy of a word-count-based program such as LIWC, it appears that this approach has several limitations for content analysis. However, if both the program developer and the end user are careful and reflexive in their consideration of theoretical assumptions, such limitations can be addressed. LIWC appears to perform better in conjunction with other, more qualitative analysis tools, and it has become clear from the experiments presented that context is paramount when measuring meaning in texts.

Acknowledgements

The author is grateful to the anonymous reviewers for their helpful comments on an earlier draft.

References

Georg W. Alpers, Andrew J. Winzelberg, Catherine Classen, Heidi Roberts, Parvati Dev, Cheryl Koopman and C. Barr Taylor. 2005. Evaluation of computerized text analysis in an Internet breast cancer support group. *Computers in Human Behavior*, 21(2):361-376.

Christopher R. Agnew, Paul A. M. Van Lange, Caryl E. Rusbult and Christopher A. Langston. 1998. Cognitive interdependence: Commitment and the mental representation of close relationships. Journal of personality and social psychology, 74(4):939-954.

Laurence Anthony. 2011. AntConc (Version 3.2.2) [Computer software]. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net.

Younggue Bae and Hongchul Lee. 2012. Sentiment analysis of Twitter audiences: Measuring the positive or negative influence of popular twitterers. *Journal of the American Society for Information Science and Technology*, 63(12):2521-2535.

Pedro P. Balage Filho, Thiago A. S. Pardo and Sandra M. Aluısio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara*.

Douglas Biber. 1988. *Variation across speech and writing.* Cambridge: CUP.

Alyssa N. Bryant, Jeung Yun Choi and Maiko Yasuno. 2003. Understanding the religious and spiritual dimensions of students' lives in the first year of college. *Journal of College Student Development*, 44(6):723-745.

Michael A. Cohn, Matthias R. Mehl and James W. Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, 15(10):687-693.

Emma Franklin. 2015. *Fresher perspectives: the transition to student life examined using LIWC and a keywords analysis*. M.Res. thesis, University of Birmingham.

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Patrick W. Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* Cambridge, MA: MIT Press.

Dan Jurafsky, Rajesh Ranganath and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638-646. Association for Computational Linguistics.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35-65.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85-88. Association for Computational Linguistics.

Matthias R. Mehl. 2006. Quantitative text analysis. In Michael Eid and Ed Diener (eds.) *Handbook of multimethod measurement in psychology*, pages 141-156.

Alexander Osherenko and Elisabeth André. 2007. Lexical affect sensing: Are affect dictionaries necessary to analyze affect?. In *Affective Computing and Intelligent Interaction*, pages 230-241. Springer Berlin Heidelberg.

Myle Ott, Claire Cardie and Jeffrey T. Hancock. 2013. Negative Deceptive Opinion Spam. In *HLT-NAACL*, pages 497-501.

Georgios Paltoglou, Stéphane Gobron, Marcin Skowron, Mike Thelwall and Daniel Thalmann. 2010. Sentiment analysis of informal textual communication in cyberspace. In *Proceedings of Engage 2010, Springer LNCS State-of-the-Art Survey*, pages 13-25.

James W. Pennebaker. 2011. *The Secret Life of Pronouns: What Our Words Say About Us.* New York: Bloomsbury.

James W. Pennebaker. 1997. Writing about emotional experiences as a therapeutic process. *Psychological science*, 8(3):162-166.

James W. Pennebaker, Roger J. Booth and Martha E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC [Computer software]. Austin, TX: LIWC.net [online]. http://liwc.net.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales and Roger J. Booth. 2007. The development and psychometric properties of LIWC2007. Austin, TX: LIWC.net.

James W. Pennebaker and Martha E. Francis. 1996. Cognitive, emotional, and language processes in disclosure. *Cognition & Emotion*, 10(6):601-626.

James W. Pennebaker, Carla J. Groom, Daniel Loew and James M. Dabbs. 2004. Testosterone as a social inhibitor: two case studies of the effect of testosterone treatment on language. *Journal of abnormal psychology*, 113(1):172.

James W. Pennebaker and Laura A. King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296-1312.

Veronica Perez-Rosas, Cristian Bologa, Mihai Burzo and Rada Mihalcea. 2014. Deception Detection Within and Across Cultures. In *Text* Mining, pages 157-175. Springer.

Stanley D. Rosenberg, Paula P. Schnurr and Thomas E. Oxman. 1990. Content analysis: A comparison of manual and computerized systems. *Journal of personality assessment*, 54(1-2):298-310.

Marcin Skowron and Georgios Paltoglou. 2011. Affect bartender - affective cues and their application in a conversational agent. In *IEEE Symposium Series on Computational Intelligence 2011, Workshop on Affective Computational Intelligence*.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24-54.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner and Isabell M. Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 178-185.

Appendix

	Finding	**Found with LIWC**	**Found with keywords**
1	More focus on collective self after the move (“we”, “our”)	X	X
2	**Social picture changes dramatically**	X	X
3	**Transition effects begin before the move date**	X	X
4	More focus on individual self after the move (“I”, “my”)	X	X
5	**Less preoccupied with media, celebrities, current affairs**	X	X
6	**More concerned with abstract ideas**	X	X
7	**More attempts to engage with/appeal to reader**	X	X
8	**More considered writing style**	X	X
9	More mentions of ‘first’, e.g. ‘our first lecture’	X	X
10	**Less general/philosophical after the move**	X	X
11	Fewer mentions of the (distant) future	X	X
12	Less focus on people, esp. other people, e.g. he, she, them	X	X
13	Preoccupation with moving before the move	X	X
14	More mentions of living arrangements	X	X
15	**Less focus on family after the move**	X	X
16	**More preoccupation with excursions, nights out**	X	X
17	**More tentative after move**	X	X
18	Less time spent on leisure activities	X	
19	Students undergo more dramatic changes than non-students	X	X
20	**More focus on food and the kitchen**	X	X
21	**Adjectives less generic**	X	X
22	**More interest in blogs and readership after the move**	X	X
23	**Appear more self-aware after the move**	X	X
24	**More mention of feelings**	X	X
25	**More comparisons and similes – describing, defining**	X	X
26	**Less preoccupation with general groups/society**	X	X
27	Fewer (rhetorical) questions	X	
28	**More advice and predictions after the move**		X
29	**More wisdoms before the move**		X
30	**Dip in ‘we’ words immediately before moving**	X	
31	**Preoccupation with flu after the move**	X	
32	**Concerned with sleep and lack thereof**	X	
33	**More mentions of favourite music, films, etc.**	X	
34	**Poetry more common after the move**	X	
35	**Fewer mentions of religious words**	X	
36	**Emphasis on what they are capable of doing after move**		X
37	**More focus on recent past; less reminiscing**		X
38	**More obligations before the move**		X

Table 3: All findings from the Franklin (2015) study regarding student changes, listed in descending order of amount of evidence to corroborate the finding. All findings in **bold** relied on examination of context to be found.

Unsupervised Topic-Specific Domain Dependency Graphs for Aspect Identification in Sentiment Analysis

Sarah Kohail
Technische Universitat Darmstadt,
Language Technology Group
Department of Computer Science
kohail@lt.informatik.tu-darmstadt.de

Abstract

We propose to model a collection of documents by means of topic-specific domain dependency graphs (DDGs). We use LDA topic modeling to detect topics underlying a mixed-domain dataset and select topically pure documents from the collection. We aggregate counts of words and their dependency relations per topic, weigh them with Tf-Idf and produce a DDG by selecting the highest-ranked words and their dependency relations. We demonstrate an implementation of the approach on the task of identifying product aspects for aspect-oriented sentiment analysis. A large corpus of Amazon reviews is used to identify product aspects by applying syntactic filtering to the DDG. Evaluation on a small set of cameras reviews demonstrate a good precision of our method. To our knowledge, this is the first method that finds product-class specific aspects in mutli-domain collections in an unsupervised fashion.

1 Introduction

Cohesion is reflected by grammatical and semantic relationships between lexical items, and links sentences together to form texts (Halliday and Hasan, 1976). These relationships contribute to the overall meaning of the text and maintain the inter-sentence and intra-sentence cohesive structure. Representations, such as graph-based have shown a potential ability to hold and understand these relationships, and facilitate knowledge extraction by enabling a variety of analysis processes (Radev and Mihalcea, 2008).

Recently, a large body of work has been devoted to applying graph or network-based methods to Natural Language Processing (NLP) problems, including, but not limited to, dependency parsing (Tzouridis and Brefeld, 2013) to semantic annotation (Nivre and Mcdonald, 2008) to text summarization (Vidal et al., 2014) and information retrieval (Blanco and Lioma, 2012). In this paper, we present a generic graph-based method and apply it to identify product aspects for sentiment analysis.

E-commerce and social media technologies have become an excellent platform for a huge number of users to share and explain their opinions online. Websites (e.g., amazon.com, flipkart.com), allow users to post and read reviews about various services and products. Such reviews are important for customers to make a purchase choice, as well as for organizations to monitor and improve their products and reputation. However, user-generated reviews are unstructured and noisy. In the past few years, there has been a significant body of work that adopts NLP tools to better understand, analyze and process arguments and opinions from various types of information in user-generated reviews. Such efforts have come to be known as sentiment analysis or opinion mining, see (Liu, 2012) for a survey.

Sentiment analysis and opinion mining have been investigated on the document level, the sentence level and the aspect level (Liu, 2012). Aspect-level sentiment analysis performs fine-grained analysis by extracting or identifying the aspects of entities and the sentiment expressed toward each extracted aspect. For example, a review of a camera is likely to discuss distinct aspects like zoom, lens, resolution, battery life, price, and memory. In exploring the problem of aspect-based sentiment analysis, we distinguish between two terms ”aspect identification” and ”aspect extraction”. Aspect extraction focuses on finding the aspects offsets in a given text reviews, while identification define the list of aspects of a certain entity.

The aim of this paper is to propose an unsupervised generic method to model a multi-domain

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 16–23,
Hissar, Bulgaria, 7-9 September 2015.

document collection by the means of domain dependency graphs (DDGs). An implementation of our method is applied to solve the aspect identification task from a large set of Amazon product reviews. The obtained graphs are used to improve the overall understanding of opinion patterns and to distinguish the most effective aspects for different product categories. Our method is completely unsupervised and needs no labeled training data or previous knowledge about the domains, and follows the Structure Discovery paradigm (Biemann, 2011). The remainder of this paper is organized as follows: Section 2 discusses related works. Section 3 describes the proposed solution. Section 4 presents and discusses our experimentation results and evaluation, followed by conclusions and future work in the last section.

2 Related Work

Graph theory has been widely used by many approaches in the field of natural language processing, text visualization and open information extraction (Koopman et al., 2012; Tzouridis and Brefeld, 2013), see (Mihalcea and Radev, 2011) for a survey. The most closely related work to our approach is (Stanovsky et al., 2014). It outlines Proposition Knowledge Graphs for information discovery. The utility of these knowledge graphs for structured queries, summarization and faceted search have been demonstrated.

In the field of sentiment analysis, graph-based approaches have been introduced to detect subjectivity (Esuli and Sebastiani, 2007; Wiebe and Mihalcea, 2006; Yu et al., 2011) or measure sentiment similarity between reviews (Goldberg and Zhu, 2006). Several methods were proposed to identify product aspects from reviews by selecting highly frequent nouns as product features (Blair-Goldensohn et al., 2008; Hu and Liu, 2004). For each detected noun, the sentiment regarding this noun is judged by its nearest adjacent adjective opinion word. However, the limitation of these methods is that many frequent noun phrases that may not represent product aspects are retrieved.

Recent research concentrates more on defining opinion patterns and relating aspects with their appropriate opinion words. Methodologies proposed in this area learn rules and templates from fully labeled data, and then use them later to detect aspects in an unlabeled dataset (Jin et al., 2009; Yu et al., 2011). Semi-supervised approaches try to reduce the amount of manual labeling by expanding a small seed set of labeled examples. Although these methods have been applied successfully in specific domains, sentiment classification is sensitive to the domain of the training data and extensive annotation for a large set of data for every single domain has to be carried out, which is not practically feasible (Vázquez and Bel, 2013).

Efforts for cross-domain sentiment analysis apply domain adaptation by limiting the set of features to those that are domain independent (Jakob and Gurevych, 2010; Li et al., 2012; Remus, 2012). An issue with these methods is that words and phrases used for expressing opinions can differ considerably from one domain to another.

3 Methodology

The purpose of this work is to advance understanding of a specific domain from mixed-domain documents by building compact directed DDGs. DDG aggregates individual dependency relations between domain-specific content words for a single topic. It gives a good visualization and summarization to a certain domain, and facilitate information and relation extraction. In this paper, we demonstrate the usage of DDGs for product aspects identification.

We summarize the methodology as follows: after preprocessing the text, we applied LDA topic modeling to discover underlying topics in a collection of textual data, and calculate a probabilistic topic distribution to select the most related phrases to each topic. POS tagging and dependency parsing were used then to select essential domain-specific phrases and content words. Finally, we build aggregate DDG per topic from the dependency parses, and use Tf-Idf and word frequency measures to weight the graph nodes and edges. A detailed discussion of our approach is given in the next section.

3.1 Dataset Preprocessing and Topic Modeling

Preprocessing includes filtering stop words, very short documents and documents with low frequency words. We perform word tokenization, and Latent Dirichlet Allocation (LDA) is then applied to extract dominant topics behind corpus of documents (Blei et al., 2003). LDA is a probabilistic graphical model that treats document as a multinomial distribution of topics, and each topic is a multinomial distribution of words. LDA is completely unsupervised and requires no human annotation, but the user has to provide the number of topics *n*. We use the implementation provided by (Phan and Nguyen, 2007). We perceive all texts belonging to one topic *i* as one document d_i, where i $\in$ $\{0, ..., n\}$. The terms ”domain” and ”topic” are used interchangeably throughout the text.

3.2 Segmentation and Preprocessing

We use the vocabulary distribution of the documents produced by LDA to find a collection of topically pure documents. We retain only documents that have a single dominating topic, which covers at least 60% of the document[1]. This step is significant to eliminate documents that contain too much noise or are too general to be characterize a specific topic. We then perform sentence segmentation[2] followed by POS tagging and collapsed dependency parsing[3] (de Marneffe et al., 2006). The output from this step is important for generating syntactic features which will be used later to filter DDGs and extract topically pure relations.

3.3 Filtering Non-Content Words

For each document d_i, collapsed dependency document is generated. It includes a set of directed typed dependency relations R_{ijk} between a head word w_{ij} and a modifier word w_{ik}. As non-content words do not contribute as much information about a specific topic, we only retain relations between content words, i.e. (common and proper) nouns, adjectives, verbs and adverbs. From this step, the work followed is done completely on collapsed dependency documents.

3.4 Term Frequency-Inverse Document Frequency (Tf-Idf)

Tf-Idf is a standard term weighting method based on their importance within a document. The core idea behind Tf-Idf is: a word *j* w_{ij} in document *i* is more relevant as a keyword for d_i if it appears many times in d_i and very few times or none in other set of documents in a corpus *D*. Tf-Idf is expressed by the following equation:

$$Tf\text{-}Idf(w_{ij}, d_i, D) = Tf(w_{ij}, d_i) \times Idf(w_{ij}, D) \qquad (1)$$

where Tf is the number of times that word *w* occurs in document d and Idf is calculated by dividing the total number of documents in a corpus, which is the number of topics *n*, by the number of documents containing the word *w* in a set of documents *D*.

Tf-Idf is calculated in three levels of granularity:

1. Word level: for each word w_{ij} in d_i, we calculated Tf-Idf using Equation 1.

2. Pair level: for each pair of words w_{ij} and w_{ik} in d_i, occurred together in a typed dependency relation R_{ijk}, we calculated Tf-Idf using the following equation:

   $$Tf\text{-}Idf(w_{ij}w_{ik}, d_i, D) = Tf(w_{ij}w_{ik}, d_i) \times Idf(w_{ij}w_{ik}, D) \qquad (2)$$

   w_{ij} and w_{ik} represents the j^{th} and k^{th} words in document *i*. Order of words w_{ij} and w_{ik} within the relation is not considered at this level.

3. Relation level: for each typed dependency relation R_{ijk} in d_i between two words w_{ij} and w_{ik}, we calculate Tf-Idf using the following equation:

   $$Tf\text{-}Idf(R_{ijk}w_{ij}w_{ik}, d_i, D) = \qquad (3)$$
   $$Tf(R_{ijk}w_{ij}w_{ik}, d_i) \times Idf(R_{ijk}w_{ij}w_{ik}, D)$$

3.5 Domain Dependency Graphs (DDGs)

DDGs are directed graph with labeled nodes and labeled edges. For each document d_i, DDG_i is constructed by aggregating individual dependency relations between domain-specific content words. DDG_i=$\{V_i$,$E_i\}$, where nodes represent words, that is V_i=$\{w_{ij}$ | $w_{ij} \in d_i$, *Tf-Idf(*w_{ij},d_i,*D)* $\geq \alpha_1$, *Tf(*w_{ij}*)* $\geq \alpha_2\}$, and edges E_i connect content words by the means of dependency relations. E_i =$\{$*(*w_{ij},w_{ik}*)* | w_{ij},$w_{ik} \in d_i$, *Tf-Idf(*w_{ij} w_{ik},d_i,*D)* $\geq \beta_1$, *Tf(*w_{ij} w_{ik}*)* $\geq \beta_2$, *Tf-Idf(*R_{ijk} w_{ij} w_{ik},d_i,*D)* $\geq \lambda_1$, *Tf(*R_{ijk} w_{ij} w_{ik}*)* $\geq \lambda_2$ $\}$.
Thresholds, α_1, α_2, β_1, β_2, λ_1, λ_2 are defined

[1] Threshold was determined in preliminary experiments

[2] Using lt.seg script from https://github.com/tudarmstadt-lt/lt.core/

[3] We use the Stanford Natural Language Processing tools http://nlp.stanford.edu/software/

by the user, and edges are labeled by the frequency and the type of dependency relation between words. Using Tf-Idf for weighting words and relations, have proven a potential ability to highlight a large set of domain-specific words and relations as will be demonstrated in the next section.

3.6 Extracting Domain Dependency Words and Relations - Application

We apply our generic approach to identify opinion phrases, and aspects of products for the use in aspect-based sentiment analysis.

Figure 1 illustrates a snapshot from DDG for a topic that captures camera reviews. We use DDGs along with Tf-Idf weighting as an important input to distinguish most related domain specific words and relation patterns. We present bellow some words examples from the camera's domain categorized by POS tags. All mentioned words are strongly related to camera domain and this proves the capability of Tf-Idf weighting in capturing potential domain specific words.

- **Adjectives:** digital, 50mm, focal, 200mm, optical, sharp, indoor, blurry, wide, prime, compact, chromatic.
- **Nouns:** lens, camera, canon, nikon, SLR, EF, shots, shutter, USM, telephoto, aperture, macro, flash, sigma, focus, pictures, zoom, tripod, powershot.
- **Verbs:** taking, focuses, capture, carry, photographing, fit, produce, cropping, adjust.

We highlight some opinion relations from Figure 1 in Table 1. The table shows dependency relation type R_{Camjk}, source word w_{Camj}, destination word w_{Camk}, relation frequency *Tf* and relation level *Tf-Idf*. We create DDGs for another 14 topics including: movies, coffee makers, electrovoice, shoes and footwear, hair products, food and baking machines, films, mp3 players, cars, TVs, mobiles, computers and perfumes. We observed that in all these graphs, opinions or relations between opinion word and opinion target, are mostly expressed with either adjectival modifier (amod) or nominal subject (nsubj). Thus, we will limit the identification of product aspects to these two dependency relations in our application.

On the basis of our analysis of DDGs and their parameters, and a list of about 6800 words positive and negative English opinion words[4], we apply a set of appropriate filters to DDG to extract opinion phrases. We filter out noun compounds relations, and words and relations below thresholds α_1, α_2, β_1, β_2, λ_1, λ_2. Either w_{ij} or w_{ik} should be in opinion words lexicon and relation which is either ”amod” or ”nsubj” is selected.

R_{Camjk}	w_{Camj}	w_{Camk}	***Tf***	***Tf-Idf***
amod	lens	fast	146	770.60
nsubj	great	lens	121	638.65
amod	picture	good	205	467.88
amod	images	sharp	116	451.45
nsubj	sharp	images	93	388.69
amod	photos	great	105	269.85
amod	picture	clear	84	241.93
nsubj	good	quality	142	50.85

Table 1: Opinion dependency relations from the camera topic.

4 Experiments

To evaluate our approach, we use an unlabeled version of Amazon dataset[5] that has been commonly used in opinion mining research (Kiritchenko et al., 2014; Tutubalina, 2015). The corpus consists of ∼35 million reviews (∼18.4 million unique reviews), about ∼2.5 million products from 28 different categories, up to March 2013. Reviews include product and user information, ratings, and a plain text review (McAuley and Leskovec, 2013).

In this work, we only use the plain text. We filter redundant reviews, reviews with less than 3 words and noisy reviews which contain smiley codes only or punctuations only, as we consider these not relevant for aspect identification. The final number of reviews we use to train the LDA model is ∼13.93 million reviews. As we mentioned in Section 3.2, we use the LDA model to select topically pure reviews. This step reduces the number of reviews to ∼1 million.

We experimentally determined a reasonable number of topics *n* to be 200, which is in line with other works using LDA for information extraction e.g. (Chambers and Jurafsky, 2011). Of the 200 topics we induced with LDA, we observed a large number of product-specific topics, as well as some mixed topics and spurious topics (Mimno et al., 2011). For this study, we proceed with selecting the 15 topics we mentioned in Section 3.6. To

[4] English Opinion Lexicon `http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon`

[5] SNAP: Web data: Amazon reviews `https://snap.stanford.edu/data/web-Amazon.html`

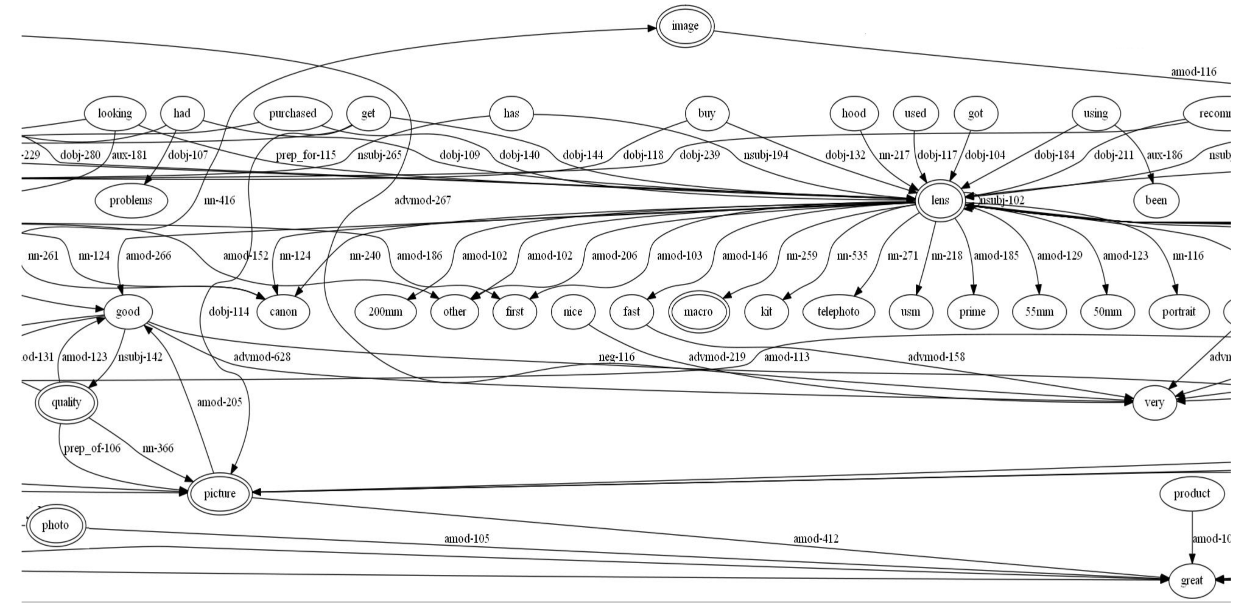


Figure 1: An excerpt from the automatically generated DDG of the camera review topic. Double lining for aspect nodes, and bold lines for connections between opinion words and aspects have been assigned manually. Only most frequent relations are shown for the purpose of presentation.

test the performance of our proposed approach, we compare our results to those obtained using DDG without Tf-Idf filtering, i.e. $\alpha_1 = \beta_1 = \lambda_1 = 0$. We evaluate the identification of aspects manually by human judgment: We order the identified relations from both Tf-Idf-based filtered DDGs (as explained in Section 3.6) and frequency-based (FB) filtered DDGs according to relation frequency. For the top 50 unique aspects, we judge whether it is an aspect of the product category or not.

Table 2 shows the experimental results for 5 different product topics. The experimental results show that Tf-Idf filtering outperforms FB filtering in terms of the number of identified aspects and it has not been worse in any case. FB ranking tend to identify general aspects such as: price, shipping, quality, value, service and company. Ranking DDGs by the means of Tf-Idf weights, gives our method the ability to detect detailed domain aspects, which is clearly evident in the cars topic in Fig. 2. The aspect identification method based on the DDG with Tf-Idf weighting identifies domain-specific aspects with an average accuracy of 53% across the five topics. When not using Tf-Idf weighting, the method achieves only an accuracy of 37%.

Our error analysis shows that most false positives by the Tf-Idf-based method consist of product domain-specific words that are not aspects. Examples from cameras domain are: fast results, great job cheap camera, excellent choice, sharp razor, perfect bag, great portrait, advanced photographer, easy c330. On the other hand, frequency-based ranking provides general noisy errors like: problem only, buy great, complaint only, time hard, addition great, drawback only, light available, room enough.

To evaluate the identified aspects coverage for the aspects extraction task from a set of reviews, we manually annotated aspects in a set of 50 cameras reviews collected randomly from Amazon. Only explicit aspects are annotated. Implicit aspects are not annotated. In most of implicit aspect expressions, adjectives and adverbs are used to describe some specific attributes of entities, for example, expensive describes price, and heavy describes weight (Liu, 2012). We compared the annotated aspects against the 33 aspects for cameras domain listed in Table 2. Out of 183 annotated aspects in the 50 reviews, 115 aspects are extracted, approximately 63%, while 38 unique failed to be extracted. Most of missed aspects are contained in cameras reviews DDG before filtering. Changing the filtering parameters can help increasing the aspects coverage but may also increase the false positive rate.

In summary, our evaluation shows a clear improvement using Tf-Idf-based filtering over the

Category / Thresholds	Method	Ext. /50	Extracted Aspects	
			Common	Difference
Camera α_1: 100, α_2:180 β_1: 2, β_2: 2 λ_1:7, λ_2:5	Tf-Idf-based	30	lens, pictures, shots, quality, images, photos, focus, light, depth, color, zoom, size, range, distortion, card, autofocus, speed.	tripod, resolution, controls, battery, mode, contrast, optics, flash, sharpness, software, screen, flexibility, distance.
	FB	20		price, value, capability.
TV α_1: 50, α_2:20 β_1: 1, β_2:1 λ_1:2, λ_2:5	Tf-Idf-based	22	cable, picture, quality, remote, setup, image.	system, audio, resolution, output, video, tuner, hdtv, quality, connection, capability, control, speakers, screen, model, component, connector.
	FB	13		price, sound, value, shipping, colors, monitor, pixels.
Computer α_1:150, α_2:50 β_1: 2, β_2:2 λ_1:2, λ_2:5	Tf-Idf-based	29	card, software, memory, adapter, performance, setup, support, camera, driver, ram, disk, space, cable.	upgrade, programs, ports, system, processor, speed, motherboard, version, machine, units, USB, slots, OS, mouse, graphics, interface.
	FB	19		price, power, value, quality, shipping, case.
Mobile α_1: 50, α_2:20 β_1: 1, β_2:1 λ_1:5, λ_2:1	Tf-Idf-based	20	sound, keyboard, screen, price, reception, quality, size, case, camera, service, software.	pictures, apps, life, interface, looks, bluetooth, battery, version, calls.
	FB	18		card, program, version, design, charger, player, value.
Cars α_1: 20, α_2:5 β_1: 2, β_2:1 λ_1:5, λ_2:1	Tf-Idf-based	32	price, performance, exhaust, wiring, plugs, installation, power, length, kit, sound, shocks, sensors, ride, instructions, parts.	work, rumble, breaks, pads, muffler, replacement, wipers, harness, connectors, idle, engine, hitch, system, unit, lights, mileage, tensioner.
	FB	23		quality, shipping, value, struts, company, service, look, room.

Table 2: Manual evaluation for aspect identification on five different domains using DDG with Tf-Idf ranking and FB ranking. It shows the number of true identified aspects out of the top 50 frequent captured relations, common identified aspects along with the difference between the two methods. The first column shows the thresholds setting. For the frequncy-based ranking method, α_1 = β_1 = λ_1 = 0.

FB baseline. This, however, is only possible for mixed-domain document collections, as Idf for a single topic is not defined.

5 Conclusion

We have introduced a new generic approach to identify the most important concepts from multi-domain document collections. Using LDA, we provided a fully unsupervised framework for extracting dominant topics behind corpus of documents, while the DDG representation maintains the inter-topic cohesiveness. Tf-Idf ensures the extraction of highly domain-specific words and relations. We demonstrate the effectiveness of the proposed approach on the task of extracting product aspects for sentiment analysis. The comparison between the DDG method and a frequency-based ranking confirms the superiority of DDG in extracting domain-specific aspects. Evaluation of DDG on a small set of cameras reviews resulted in a precision of ∼63%. This is the first approach, to our knowledge, for extracting product aspects from mixed-domain dataset, without the use of an external knowledge base or a training dataset.

In the future, we hope to advance our work by using DDGs to applying more advanced ranking and filtering techniques to DDGs such as centrality (Newman, 2010) or PageRank (Brin and Page, 1998) for node ranking. Collecting similarities to the existing list of aspects and grouping aspects using techniques from distributional semantics would improve the overall recall.

Acknowledgments

This research was supported by the DAAD. The authors sincere gratitude goes to Chris Biemann for his full supervision. Also, the author would like to thank the members of Language Technology group in TU Darmstadt and the anonymous reviewers who greatly refined the drafts of this paper.

References

Chris Biemann. 2011. *Structure Discovery in Natural Language*. G. Hirst, E. Hovy and M. Johnson (Series Eds.): Theory and Applications of Natural Language Processing. Springer, Heidelberg Dordrecht London New York.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George A Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews. In *WWW Workshop on NLP in the Information Explosion Era*, pages 14–23, Beijing, China.

Roi Blanco and Christina Lioma. 2012. Graph-based term weighting for information retrieval. *Information retrieval*, 15(1):54–92.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Nathanael Chambers and Dan Jurafsky. 2011. Template-based Information Extraction Without the Templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 976–986, Portland, Oregon.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, volume 6, pages 449–454, Genoa, Italy.

Andrea Esuli and Fabrizio Sebastiani. 2007. Pageranking WordNet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics ACL-07*, volume 7, pages 442–431, Prague, Czech Republic.

Andrew B Goldberg and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, New York, NY, USA.

Michael AK Halliday and Ruqaiya Hasan. 1976. Cohesion in English. *Longman's, London*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, New York, NY, USA.

Niklas Jakob and Iryna Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045, Cambridge, Massachusetts.

Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. A Novel Lexicalized HMM-based Learning Framework for Web Opinion mining. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 465–472, Montreal, Quebec, Canada.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland.

Bevan Koopman, Guido Zuccon, Peter Bruza, Laurianne Sitbon, and Michael Lawley. 2012. Graph-based concept weighting for medical information retrieval. In *Proceedings of the Seventeenth Australasian Document Computing Symposium*, pages 80–87, Dunedin, New Zealand.

Fangtao Li, Sinno Jialin Pan, Ou Jin, Qiang Yang, and Xiaoyan Zhu. 2012. Cross-domain co-extraction of sentiment and topic lexicons. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 410–419, Jeju Island, Korea.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, Hong Kong, China.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based natural language processing and information retrieval*. Cambridge University Press.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Edinburgh, United Kingdom.

Mark Newman. 2010. *Networks: an introduction*. Oxford University Press.

Joakim Nivre and Ryan Mcdonald. 2008. Integrating Graph-Based and Transition-Based Dependency Parsers. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT*, pages 950–958, Columbus, Ohio.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA). `http://gibbslda.sourceforge.net`.

Dragomir R Radev and Rada Mihalcea. 2008. Networks and natural language processing. *AI magazine*, 29(3):16–28.

Robert Remus. 2012. Domain adaptation using domain similarity-and domain complexity-based instance selection for cross-domain sentiment analysis. In *IEEE 12th International Conference on Data Mining Workshops (ICDMW 2012)*, pages 717–723, Brussels, Belgium.

Gabriel Stanovsky, Omer Levy, and Ido Dagan. 2014. Proposition Knowledge Graphs. In *Proceedings of the First AHA!-Workshop on Information Discovery in Text*, pages 19–24, Dublin, Ireland.

Elena Tutubalina. 2015. Target-Based Topic Model for Problem Phrase Extraction. In *Advances in Information Retrieval*, pages 271–277. Springer.

Emmanouil Tzouridis and Ulf Brefeld. 2013. Learning the Shortest Path for Text Summarisation. In *The Fourth International Workshop on Mining Ubiquitous and Social Environments*, pages 45–57, Prague, Czech Republic.

Silvia Vázquez and Núria Bel. 2013. A classification of adjectives for polarity lexicons enhancement. *arXiv preprint arXiv:1303.1931*.

Juan C Vidal, Manuel Lama, Estefanía Otero-García, and Alberto Bugarín. 2014. Graph-based semantic annotation for enriching educational content with linked data. *Knowledge-Based Systems*, 55:29–42.

Janyce Wiebe and Rada Mihalcea. 2006. Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1065–1072, Sydney, Australia.

Jianxing Yu, Zheng-Jun Zha, Meng Wang, and Tat-Seng Chua. 2011. Aspect ranking: identifying important product aspects from online consumer reviews. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1496–1505, Portland, Oregon.

Unsupervised learning of agglutinated morphology using nested Pitman-Yor process based morpheme induction algorithm

Arun Kumar
Universitat Oberta Catalonia, UPC
arunsocs@gmail.com

Abstract

In this paper we describe a method to morphologically segment highly agglutinating and inflectional languages from Dravidian family. We use nested Pitman-Yor process to segment long agglutinated words into their basic components, and use a corpus based morpheme induction algorithm to perform morpheme segmentation. We test our method in two languages, Malayalam and Kannada and compare the results with Morfessor Categories-MAP.

1 Introduction

Morphological processing is an important task for natural language processing systems, such as information retrieval systems. In the case of languages with agglutinated and rich morphology, such as Dravidian family of languages, morphological processing is more important because one word can actually be the combination of several others, each with a number of morphological/flexive markers. Properly identifying morphemes in agglutinated words is essential for tasks such as information retrieval and machine translation.

Consider the following example from Malayalam, a language from south Dravidian family having 38 millions of native speakers and one of the classical languages of India. A word in Malayalam (പുഴകളയിരുന്നു, pul̲akaḷayirunnu, there were rivers), here root word is (പുഴ, pul̲a, river) is inflected with plural marker (കൾ, kal, Plural marker) and it also contains verb phrase(ആയിരുന്നു, ayirunnu were) all of them are joined together with orthographic changes. It is possible to have orthographic changes when words are combined, because of morpho-phonemic change called sandhi, which makes the task of segmenting Dravidian languages challenging. Dravidian languages are agglutinated like Turkish and inflected like Finnish. Other than agglutination and inflection, Orthographic changes in morpheme boundaries occurs due to sandhi changes and alpha syllabic writing system. In this case the job of a morphological analyzer is to segment the large word sequence (പുഴകളയിരുന്നു, pul̲akaḷayirunnu, there were rivers) into(പുഴ, pul̲a കൾ, kal, ആയി, ayi, ഉന്നു, unnu), which are the constituent morphemes. In the above word phrase orthography of constituent morphemes are different when they combined to in the form a word due to alpha-syllabic script that capture phonological changes. This property makes morphological processing of this languages challenging. As morpheme boundaries are marked at syllabic level, morpheme boundaries can occur inside ligatures an digraphs. In this paper we are developing a non parametric Bayesian models based on nested Pitman-Yor process on syllable level to segment long words into individual components and learn their morphological segmentation.

Dravidian family of languages are least resourced so we use corpora created from Wikipedia for conducting the experiments. We define a nested Pitman-Yor process based model for segmentation of agglutinated long sequence of words and defined model inferred using a parallel blocked Gibbs sampling algorithm. It is a generative approach in which we consider syllables are the basic units that are combined in context (agglutination) to form words. Once the algorithm achieves the segmentation on corpus created from Wikipedia, we use a heuristic search based algorithms to achieve final morphological segmentation. We test our algorithm pipeline in the case of two highly agglutinated and inflected lan-

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 24–29,
Hissar, Bulgaria, 7-9 September 2015.

guages, Malayalam and Kannada from Dravidian family. As the gold standard segmentation is not available for evaluation, we created a gold standard segmentation file for both languages and evaluate the results. We manually analyze the errors in morphological segmentation to get the idea of errors that are produced by them system and to improve the system performance in further studies. In section 2 we describe previous work Bayesian non-parametric and morphological processing of agglutinating languages. In section 3 we describe Pitman-Yor models, and Section 4 describes the used algorithm for morphological segmentation. Sections 5 and 6 present the results and error analysis, and finally, section 7 presents the conclusions and future work of our research.

2 Related Work

In this section we describe related works carried out on Bayesian non-parametric models to learn morphology of languages. Research works in unsupervised learning of morphology are also relevant. Hammarström and Borin (Hammarström and Borin, 2011) provide a detailed survey of the topic. Morfessor (Creutz and Lagus, 2002; Creutz and others, 2006; Creutz et al., 2007) based on Minimum Description Length principle is the reference model for highly inflecting languages, such as Finnish. Goldwater et al. (Goldwater et al., 2009) introduce a word segmentation model based on Dirichlet Process mixture to model words and their contextual dependencies. They test their method on phonetic scripts of child speech. Following this line of research, Naradowsky & Goldwater (Naradowsky and Goldwater, 2009) incorporated English spelling rules to the morphological model to achieve better results for English phonetic script segmentation. Following these studies, Teh (Teh, 2006) introduced a Bayesian language model based on Pitman-Yor process and a new sampling procedure for the model. Lee et al. (Lee et al., 2011) modeled syntactic context to achieve better morphological segmentation. Dreyer & Eisner (Dreyer and Eisner, 2011) identified morphological paradigms using Dirichlet Process Mixture models and seed paradigms. Can and Manandhar (Can and Manandhar, 2012) clustered morphological paradigms using Hierarchical Dirichlet Process models, and Sirts & Goldwater (Sirts and Goldwater, 2013) used adapter grammar to achieve morphological segmentation. Nested Pitman-Yor process is an extension of above Dirichlet process, used to produce word segmentation of languages, such as Japanese (Mochihashi et al., 2009) and creation of language models for speech recognition (Mousa et al., 2013). These works are also relevant in the case of Bayesian non parametric models for learning morphology.

In the case of the Dravidian languages, unsupervised techniques are rarely applied. For the larger languages of the family (Telugu, Tamil, Kannada and Malayalam) there are studies that use supervised techniques. Those studies in the case of Malayalam are the following: Vasudevan & Bhattacharya (N and Bhattacharyya, 2013) propose a stemmer for Indian languages, such as Hindi, Marathi and Malayalam based on suffix lists. Idicula & David (Idicula and David, 2007) present a morphological analyzer for Malayalam based on Finite state Transducers and inflectional rules.

3 Pitman-Yor Process language model

Pitman-Yor process (Pitman, 2002) a generalization of Dirichlet process and it is a stochastic process. Goldwater et al. (Goldwater et al., 2009) and Teh (Teh, 2006) use it for language modeling. It is represented as:

$$G \sim PY(G_0, d, \theta)$$

The stochastic process generates a discrete probability distribution G similar to another given distribution $G0$. G_0 is called base measure, d is a discount factor and θ is a variable that controls similarity between both distributions G_0 and G.

A unigram language model can be expressed as a Pitman-Yor process as:

$$G_1 = p(w) \quad \forall w \in L$$

where w ranges over all words in the lexicon (L).

In the case of a bigram distribution, we have

$$G_2 = p(w|v) \quad \forall v, w \in L$$

For frequent words G_1 will be similar to G_2, so we can compute G_2 using G_1 as a base measure:

$$G_2 \sim PY(G_1, d, \theta)$$

Similarly it is possible to compute also trigram models. As this model has no analytic form the model described is represented in the form of Chinese Restaurant Process (CRP) (Aldous, 1985). Chinese Restaurant Process is an infinite large restaurant with infinitely many tables and capacity of many customers. At first the restaurant is empty, then the first customer enters and sit at an empty table. Next customer sit a new table, based on a concentration parameter or sit to already occupied table probability proportional to number of customers sitting there.

n - gram probability computed in CRP representation. Words are customers that are sitting in various tables. Tables in the restaurants are context of the words. Context of the word is length of the suffix in all earlier occurrences. So in this representation, each n-gram context h is a table and customers are n-gram counts seated over tables $1 \cdots t_{hw}$. The seat assignation to customers is constructed choosing a table k for each $c(w|h)$ (count of w given the context h) is the n- gram count and its probability is proportional to

$$p(c(w|h)) \propto \begin{cases} c_{hwk} - d, & k = (1, \cdots t_{hk}) \\ \theta + d \cdot t_h & (k = new) \end{cases}$$

where c_{hwk} is the number of customers seated in the table k and t_h is the total number of table in h. When the $k = new$, the t_h is incremented. As a result the n-gram probability can be computed as:

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + dt_h}{\theta + c(h)} p(w|h')$$

where θ and d are the hyper parameters to be learned from data. Those parameters are inferred from the data (unsegmented corpus) and assuming that posterior probability of the variable are from Beta or Gamma distribution.

Inference on the model is done using adding and removing customers to the table t_w in the way d and θ are optimized using MCMC. For more details, refer to (Teh, 2006)

3.1 Nested Pitman-Yor process

Nested Pitman-Yor Process is a hierarchical process in which the base measure G_0 is replaced with another Pitman-Yor process. In our model base measure G_0 is replaced by a Pitman-Yor process of syllable n-grams. Then the base measure becomes:

$$G(w) = p(s_1 \cdots s_k) = \prod_{i=1}^{k} (s_i | s_{i-n+1} \cdots s_{i-1})$$

The above process can be consider as Hierarchical model, where two levels exist one is the word model and another is syllable model. We consider our syllable model as uni gram language model. For the inference it is represented in the form of a nested CRP in which a word model is connected to syllable model. In this set-up, a word w is generated from a base measure and the base measure is a Pitman - Yor process of syllables. For the inference on the particular model, we use a parallel blocked Gibbs sampler. Considering the syllables are the basic characters that joined to form words sentences. More details of sampling procedure can be found in (Neubig, 2014).

4 Morpheme identification and verification algorithm

After inference on the defined model, we apply a morpheme identification and verification algorithm to the acquired root words and morphemes. Our method is similar to that of Dasgupta & Ng (Dasgupta and Ng, 2007).

Our morpheme identification algorithm has two major parts. The first part of the algorithm is to identify a list of possible affixes for morpheme induction and composite suffixes. The list of possible affixes is extracted from the segmented corpus in following way: We assume that a word $\alpha\beta$ is concatenation of α and β, If we find both α and $\alpha\beta$ in the counter (we keep a counter of words from segmented corpus according to their frequencies) we extract β to the list of suffixes. Similarly if we find character sequence in $\alpha\beta$ and β in the counter, we list the α in the list of prefixes. But the problem with this technique is that it can create a large number of invalid suffixes and prefixes. To reduce this problem we rank the affixes based on their frequencies with different character sequences. Only top affixes that have got higher ranks are selected for induction purposes.

The second part of the algorithm aims to identify composite suffixes. As the Dravidian language family is highly inflectional large number of composite affixes are present in the vocabulary. For example in Malayalam, (ആളുകളുടെ, āḷukaḷuṭe, belongs to men) has a composite suffix (കളുടെ , kaluṭe) formed by suffixes (കൾ, kal ഉടെ, uṭe). We remove these composite suffixes from list of suffixes, otherwise it can lead to under segmentation. The third step of our morpheme identification algorithm is to identify possible roots. We take a word w from the counter and then we compose it with suffixes in the counter table. Thus, if $x + w$ (where x is an induced prefix) or $w + y$ (where x is an induced suffix) is present in the corpus, we consider w as a root and it is added to the root list. This procedure is continued until we get root, prefix and suffix lists. Using the proposed list of roots, prefixes and suffixes overall corpus is segmented to morphemes.

5 Data and Experiments

To validate our model and algorithm, we tested our algorithm on Malayalam and Kannada corpus. As Malayalam and Kannada are least resourced languages, we used a corpus crawled from Wikipedia containing 10 million words both languages, which are manually processed. As a first step of our experiments, we converted the Unicode encoded file to corresponding ISO romanized form for internal processing. We create word list of 10 million words annd add a space between characters, For example, A Kannada word (ವಿದ್ಯಾರ್ಥಿ , Vidyārthi, student) is represented as V i d y ā r t h i and it converted into constituent syllables.

Second step of the experiment consists of applying our nested Pitman-Yor model and inference algorithm to the data. For this the data is fed to the sampling algorithm for 100 iterations. Depending on the number of tokens, time taken for convergence varies. Our algorithm took 3 hours to converge in a machine with a 4-core processor with four threads in execution.

Next step is to apply our morpheme identification and evaluation algorithm to induce morpheme. Once the process is completed the system produces morphological segmentation of input words. For evaluation, we manually segmented 10,000 words of Malayalam and Kannada. The segmentation in the gold standards as follows (മനുഷ്യൻറെ,, manuṣyanr̲e, of human) The segmentation is (മനുഷ്യൻ ,manuṣyan ഇൻറെ , inr̲e Genitive case marker). We measured precision (P), recall (R) and F-measure (F) of predicted morpheme boundaries. We used programs provided by morpho-challenge (Virpioja et al., 2011) team for evaluation.

In order to get a comparison result, we train Morfessor Categories-MAP 0.9.2 [1] with same 10 million words for 10 Epoch and create the model. Using the model produced we segment the gold standard file and apply evaluation algorithm.

Results of the experiments shown in Table1

Table 1: Results compared to Morfessor-MAP

Method	Kannada			Malayalam		
	P	R	F	P	R	F
Morfessor- MAP	48.1	60.4	53.5	47.3	60.0	52.9
NPY	66.8	58.0	62.1	60.3	59.6	59.9

6 Error Analysis

We analyzed the results of experiments to get an insight errors that need to be solved in future research. We are listing the errors that are produced by our algorithms and Morfessor-MAP. In the case of our algorithm, it has two major steps one is to identify accurate word boundaries and other is to find accurate morpheme boundaries.

- Morfessor and our system fail to identify character combinations which need to considered as single character so it segmented digraphs and ligatures. In the case of our system it as we use a internal notation it did not segment the digraphs and ligatures.

- In the case of loaned root words, both systems fails to identify the morphemes.

[1] http://www.cis.hut.fi/projects/morpho/morfessorcatmapdownloadf

- Our system is able to identify morpheme boundaries where morpho-phonemic occurs. In the case of Morfessor-MAP, it fails to identify morpheme boundaries if there is a morpho-phonemic change and it consider zero-width joiner of Unicode as morpheme boundary.

- Our algorithms is able to identify orthographic changes that happening in the morpheme boundaries during sandhi changes but Morfessor-MAP fails. For example, a Malayalam word (മരങ്ങൾ, maraṅṅaḷ, trees) our system segment it to (മരം,maraṁ) and (ങ്ങൾ., ṅṅaḷ).

7 Conclusions and future research

We presented a method to segment words into morphemes using nested Pitman-Yor process for highly agglutinating and least resourced language such as Malayalam and Kannada. Our morphology learning system segmented complex morpheme sequences and it produce results that outperform state of the art systems. In future research, we focus on morphological processing of other languages in Dravidian family and we also focus on more richer models

Acknowledgments

References

David J Aldous. 1985. Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII—1983, pages 1–198. Springer Berlin Heidelberg.

Burcu Can and Suresh Manandhar. 2012. Probabilistic hierarchical clustering of morphological paradigms. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 654–663. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6, pages 21–30. Association for Computational Linguistics.

Mathias Creutz et al. 2006. Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition. Helsinki University of Technology.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. ACM Transactions on Speech and Language Processing (TSLP), 5(1):3.

Sajib Dasgupta and Vincent Ng. 2007. High-performance, language-independent morphological segmentation. In HLT-NAACL, pages 155–163. Citeseer.

Markus Dreyer and Jason Eisner. 2011. Discovering morphological paradigms from plain text using a dirichlet process mixture model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 616–627. Association for Computational Linguistics.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1):21–54.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. Computational Linguistics, 37(2):309–350.

Sumam Mary Idicula and Peter S David. 2007. A morphological processor for malayalam language. South Asia Research, 27(2):173–186.

Yoong Keok Lee, Aria Haghighi, and Regina Barzilay. 2011. Modeling syntactic context improves morphological segmentation. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 1–9. Association for Computational Linguistics.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1, pages 100–108. Association for Computational Linguistics.

Amr El-Desoky Mousa, M Ali Basha Shaik, Ralf Schlüter, and Hermann Ney. 2013. Morpheme level hierarchical pitman-yor class-based language models for lvcsr of morphologically rich languages. In INTERSPEECH, pages 3409–3413. Citeseer.

Vasudevan N and Pushpak Bhattacharyya. 2013. Little by little: Semi supervised stemming through stem set minimization. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, pages 774–780, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Jason Naradowsky and Sharon Goldwater. 2009. Improving morphology induction by learning spelling rules. In IJCAI, pages 1531–1536.

Graham Neubig. 2014. Simple, correct parallelization for blocked gibbs sampling. In Technical Report, November.

Jim Pitman. 2002. Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using adaptor grammars. TACL, 1:255–266.

Yee Whye Teh. 2006. A hierarchical bayesian language model based on pitman-yor processes. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pages 985–992, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. Traitement Automatique des Langues, 52(2):45–90.

Easy-read Documents as a Gold Standard for Evaluation of Text Simplification Output

Victoria Yaneva
Research Group in Computational Linguistics
University of Wolverhampton
v.yaneva@wlv.ac.uk

Abstract

This paper presents initial research into the use of easy-read articles written for people with cognitive disabilities as a gold standard for the evaluation of the output of text simplification systems. We investigate the compliance of the easy-read documents available on the Web with guidelines for development of easy-read material, as well as their suitability as a gold standard for simple documents for two types of populations in particular: adult readers with autism and readers with mild intellectual disability (MID). The results indicate an overall good level of compliance with the guidelines and suggest that easy-read documents are a suitable resource for evaluation of accessible documents produced for adults with autism or MID.

1 Introduction

Common reading materials on the Internet, in newspapers or in textbooks are generally understood by a large part of the population. However, there are readers with disabilities such as intellectual disability, autism, aphasia or dementia, among others, who struggle to comprehend most of these written materials.

To ensure the constitutional right of all individuals to have access to information (WHO, 2011), there is a campaign for the production of "easy-read" documents, which are accessible documents produced by humans, following a set of guidelines for accessible writing, such as the 'Make It Simple' guidelines (Freyhoff, 1998) or 'Guidelines for Easy-to-read Materials' (Nomura et al., 2010). The comprehensibility of the easy-read documents is also ensured by the inclusion of images to illustrate the main ideas in the text, and by the evaluation of these documents on a focus group of disabled people. While many governmental and healthcare organisations within the UK and the USA are required by law to produce accessible versions of important documents (Equality Act 2010), and many charity organisations produce such documents too, their writing and evaluation is time-consuming and expensive.

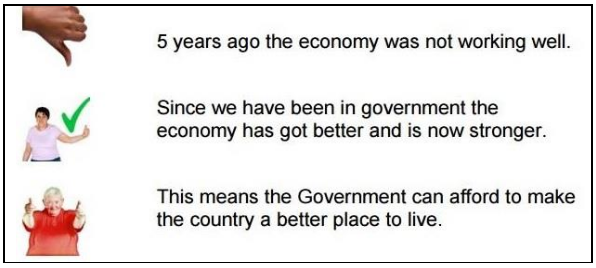
Another way to make text documents accessible for disabled readers is to convert them using automatic text simplification. Text simplification (TS) is a process which aims to enhance the understandability of a text by performing different linguistic transformations without changing its original meaning (Max, 2006). While automatic TS is promising in terms of time and financial cost, current TS systems are still not advanced enough to replace humans in the production of accessible documents. This problem is partially due to the scarcity of corpora of original and accessible texts with aligned sentences (parallel corpora) on which to train TS systems, as well as the issue of deciding which texts are simple enough for particular groups of disabled readers to be used as a gold standard for the evaluation of the TS output.

This paper describes initial research into the question of whether human-produced easy-read versions of documents could be used as a gold standard for accessible writing for particular user groups, such as readers with autism or intellectual disability. The compilation of such a corpus has now become feasible due to the already large number of existing easy-read documents produced between the early 2000s and now. Thus, for example, the original and easy-read versions of the UK political parties' manifestos from the 2015 elections (Figure 1 and Figure 2) illustrate the progress of the easy-read campaigns in adapting documents from various genres and domains:

> Five years ago, Britain was on the brink. As the outgoing Labour Treasury Minister put it with brutal candour, 'there is no money'. Since then, we have turned things around.

Figure 1: Conservative Party Manifesto (2015)

Proceedings of the Student Research Workshop associated with RANLP 2015, pages 30–36,
Hissar, Bulgaria, 7-9 September 2015.



5 years ago the economy was not working well.

Since we have been in government the economy has got better and is now stronger.

This means the Government can afford to make the country a better place to live.

Figure 2: Conservative Party Manifesto Easy-Read Version (2015)

However, before accepting such easy-read documents as the epitome of how an accessible text should look, we should keep in mind the variety of sources they come from, some of which may not be reputable. The current paper investigates two main aspects of easy-read documents on the Web: 1) their compliance with the guidelines according to which they were produced and 2) their suitability for particular target populations such as readers with autism or intellectual disability. In particular, we investigate the following questions:

- Do the easy-read documents available on the Web comply with the rules outlined in the guidelines for creation of easy-read documents?
- Are the easy-read documents available on the Web simple enough (or too simple?) for TS target groups such as people with mild intellectual disability or autism?

The first question is important because no official data for the evaluation of these documents has been published. We aim to assess compliance through, firstly, assigning linguistic features to the rules outlined in the guidelines for production of easy-read material and then analysing a sample of easy-read documents based on these linguistic features (Section 3). This analysis of compliance is going to cast light on the question of whether the majority of the easy-read material coming from a variety of sources on the Web could actually be regarded as such.

The second question of suitability is no less important due to the heterogeneity of conditions entailing reading difficulties. This means that a text at a certain readability level may be suitable for adults with mild intellectual disability (MID), but at the same time it could be too simplistic for adults with autism or too challenging for children with MID. This is an argument against the "one-size-fits-all" approach in creating accessible documents, where no distinction is made between the different levels of ability in cognitively disabled people, in addition to their individual differences. In the context of this research we define suitability as the appropriate level of difficulty of texts for particular target populations. We investigate the suitability of easy-read documents for adults with autism and MID by comparing them with corpora developed for and evaluated on such readers based on 13 relevant linguistic features (Section 4).

2 Related Work

As mentioned in Section 1, automatic TS systems make use of monolingual corpora, where the text pairs could be an original article and its simplified version (parallel corpus), or two articles with different complexity levels collected based on similar criteria (e.g. topic or timespan) (comparable corpus).

2.1 Existing Corpora

For English there are several comparable corpora, which have been used in TS tasks. Simple English Wikipedia[1] together with English Wikipedia[2] comprise probably the largest resource used in automatic TS. However, the accessibility of articles in Simple Wikipedia has been disputed, with researchers appealing for "the community to drop it as the standard benchmark set for simplification" due to its many drawbacks (Xu et al., 2015). For example, Stajner et al. (2012) compare the corpus to articles from the genres of News, Health and Fiction on the basis of 4 readability formulae and 16 linguistically motivated features and find that the articles in Simple Wikipedia are more complex than the ones in the Fiction genre (Stajner et al., 2012).

Other corpora used for TS include the relatively small EncBrit (Barzilay and Elhadad, 2003), consisting of 20 articles from Encyclopedia Britannica and their manually simplified versions for children from Britannica Elementary. Due to its small size, this corpus has been used as a test set only. WeeklyReader (Allen, 2009) and Literacyworks (Peterson and Ostenforf, 2007) also have manually simplified versions for language learners for, respectively, 100 and 104 of their articles. A larger and more recent TS resource is the parallel Newsella corpus (Xu et al., 2015), which consists of 1,130 news articles, re-written for children at 4 different grade levels.

[1] http://simple.wikipedia.org/wiki/Main Page
[2] https://en.wikipedia.org/wiki/English_Wikipedia

Currently there are only two parallel corpora for English, which have been specifically compiled for people with disabilities. The FIRST corpus consists of 25 texts and their simplified versions for people with autism (Jordanova et al, 2013). While the simplification was performed by experts working with autistic people who followed autism-specific simplification guidelines (Martos et al., 2012), the corpus was never actually evaluated by autistic readers. The other corpus, compiled by Feng et al. (2009), is called LocalNews and is comprised of 11 newspaper articles and their simplified versions for adults with mild intellectual disability (ID). Unlike FIRST, LocalNews has been evaluated by 20 adults with mild ID.

2.2 Strengths and Limitations of the Existing Corpora

A great advantage of using parallel corpora such as those mentioned above is that the original and simplified sentences are aligned, which allows automatic learning of simplification rules, hence the use of the corpora not only for evaluation but also for development of TS systems. However, currently there is no information as to whether manual simplification done with the primary objective of producing aligned corpora is of a similar quality to accessible documents produced with the reader in mind (e.g. easy-read documents). At present, this question remains an avenue for future research into the quality of resources used in TS. In addition, few of these corpora have actually been evaluated on relevant reader groups and in some cases (e.g. FIRST) the sole fact that the simplification has been done by experts is used to ensure the quality of the resource. Finally, a drawback of all corpora mentioned above is the fact that they all come from one source only (e.g. Wikipedia or Encyclopedia Britannica) and are genre-specific (encyclopedic articles, newspaper articles in the case of LocalNews, newspaper and informational articles in the case of FIRST).

Compiling a corpus of easy-read documents has the potential to overcome the issues related to source and genre, because: 1) easy-read texts come from a variety of sources (the credibility of some of them being uncertain, which is why we are first going to assess the compliance of these documents collected from the Web with their production guidelines) and 2) they cover a wider variety of genres such as newspaper articles, general informational articles, healthcare, politics, literature, fun facts, etc. (Section 3). Finally, easy-read documents are widely available and do not require the time-consuming rewriting of original articles.

3 Assessing Compliance of Easy-Read Documents On the Web

As an initial step towards the development of a corpus of easy-read documents, we have first collected a sample of 150 easy-read documents in order to assess their compliance with the guidelines for their production.

3.1 Collecting a Sample of Easy-Read Documents

The sample of 150 documents consists of 78,324 words and 12,692 sentences in total. The average number of sentences per document was 84.05 with standard deviation (SD) of 98.7 and average sentence length in words 6.3 (SD = 2.17). The average number of words per document was 518.7 (SD = 624.18). When collecting this sample, we have tried to make it a balanced representation of sources and genres. The documents included in the sample were obtained from various UK and US charity organisation websites (38 documents), government departments (26 documents), healthcare services (32 documents), as well as demos of adapted books from educational websites (3 documents) and online news websites for people with disabilities (50 documents). All documents were written in English. The topics of the documents were highly dependent on their sources and thus they encompass healthcare, news, literature, politics, policies, and general information for everyday life, which is typically provided by the charity organisations (e.g. how to shop for healthy food or how to make a doctor's appointment). In Section 3.2 we analyse some characteristics of this sample, relevant to the initial guidelines for creating easy-read texts.

3.2 Linguistic Features

There are various guidelines for creating easy-read documents (Freyhoff, 1998; Nomura, Nielsen and Tronbacke, 2010), with some charity organisations creating manuals of their own. In this paper we focus on the linguistic aspect of these documents by summarising the main points of the Make It Simple guidelines (Freyhoff, 1998), as a well-known resource for producing easy-read documents, and by analysing the easy-read sample through identifying and measuring 13 features relevant to the postulates of these guidelines. Column 1 in Table 1 lists the main recommendations of the writing guidelines, matched with corresponding linguistic features used in our analysis reflecting these recommendations (column 2). Column 3 gives the scores obtained for these features for our sample of 150 documents. The features were obtained with the Coh-Metrix 3.0 system (McNamara, 2013).

Writing Rules	Linguistic Features	Score	SD
Use short sentences	Average Sentence Length in Words	6.3	2.17
Use short words of everyday spoken language	Average Word Length in Syllables	1.44	0.12
	Word Frequency	2.43	0.2
	Age of Acquisition	317.4	35.7
	Familiarity	580.8	7.58
Use active verbs	Agentless Passive Voice Density	7.53	8.36
Use positive language	Negation Density	9.16	8.66
Use many personal words	1st Person Singular Pronoun Incidence	5.34	19.6
	2nd Person Pronoun Incidence	34.24	39.5
Avoid abstract concepts	Imagability	419.8	29.4
	Concreteness	388.9	33.4
Use simple language	Flesch Reading Ease	78.84	10.9
	Flesch-Kincaid Grade Level	3.83	1.75

Table 1: Writing rules (Freyhoff, 1998) and their corresponding linguistic features

As can be seen from Table 1, we have used the **average sentence length in words** feature as a straightforward measure of the *Length of the Sentences*. The use of *Short Words of Everyday Language* we measure through 4 indices: **word length, word frequency, age of acquisition and familiarity**, the latter two being based on norms from the MRC psycholinguistic database (Gilhooly and Logie, 1980) incorporated into the Coh-Metrix 3.0 package (McNamara, 2013). The MRC database is based on human ratings, where a word is assigned low AOA index if most people have rated it as acquired early in childhood, e.g. words such as *milk* or *pony* have a score of 202 and words such as *dogma* or *matrix* have a score of 700. The familiarity index goes into the opposite direction: a high score means that the word is very familiar for a large part of the population sample. By comparison, the familiarity of the word *milk* has received a score of 588, while *dogma* is 328.

The use of *Active Verbs* and *Positive Language* has been measured through counting the **number of passive voice and negative constructions** respectively, so the lower the scores of these indices are, the higher the readability. Use of *Personal Words* is defined in the guidelines as: "Address your readers in a direct and personal form". To account for this aspect we have included indices such as **first person and second person pronoun incidence**. *Abstractness* is measured through **imagability** and **word concreteness** indices, which aim to identify words that evoke mental images and are thus easier to process, based on human ratings from (Gilhooly and Logie, 1980). Finally, the general *Simplicity of Language* is measured through two widely-used readability formulae: **Flesch Reading Ease** where 0=very difficult and 100=very easy (Flesch, 1948) and **Flesch-Kincaid Grade Level**, where 0 = very easy and 12=very difficult (Kincaid et al., 1975).

Other rules in the guidelines, which were not evaluated in this experiment due to lack of relevant linguistic indices, were: *Use Practical Examples, Address the Readers in a Respectful Form, Cover Only One Idea per Sentence, Do Not Assume Previous Knowledge, Use Words Consistently, Do Not Use the Subjunctive Tense* and *Be Careful with Metaphors and Figurative Language.*

3.3 Results

The results indicate that, indeed, the documents in the sample used fairly short sentences of 6.3 words on average, as well as short words of 1.44 syllables on average. Most of the words have also been acquired early in childhood (**AOA** = 317.4) and are highly familiar (**familiarity index** = 580.82). **Imagability** (419.78) and **concreteness** (388.87) are also high, meaning that most of the words were not abstract. Overall, we can conclude that the lexical component of the sample complies with the requirements of the guidelines. We can also observe very few uses of **passive voice** (7.53) or **negation** (9.16) and a very high **second person pronoun incidence** (34.24), showing that the reader has often been addressed directly.

The **Flesch** and **Flesch-Kincaid** formulae also demonstrate a good level of readability of the texts. The Flesch formula has an average score of

78.84 for the sample, where a score of 0 stands for “very difficult” and a score of 100 stands for “very easy”; the Flesch-Kincaid Grade Level goes in the opposite direction (the lower the score, the easier the text) and gave an average value of 3.83 for our sample.

Even though all measures indicate a very good level of accessibility of the documents, the SD measures vary greatly, which means that some of the documents score very highly in some of the measures, while others had very low scores. To investigate this further, we ranked all texts based on the scores of the Flesch Reading Ease formula. Focusing on the lower quartile, we identified 11 texts from miscellaneous sources, the Flesch readability of which was under the recommended threshold of 65 for documents written in plain English (Flesch, 1948), with some of them going as low as 43.1 or 48.77. The Flesch Reading Ease measure was consistent with the rest of the measures in identifying these 11 texts as deviant from the other 139 ones and thus we regard these as easy-read documents with lower compliance to the guidelines and thus with potentially lower accessibility.

As a whole, the results indicate that the selected sample of accessible texts complies with the standard set out in the easy-read guidelines. Only 7.33% of the texts (11 documents) showed readability under the threshold for what could be considered an accessible document. However, it is known that readability indices are an approximation only and do not account for all aspects of the text and reader interaction (DuBay, 2004). Thus, we can conclude that easy-read documents randomly selected from various domains on the Web overall comply with the rules in the easy-read guidelines.

In the next section we compare the sample of 150 documents to other corpora, which have previously been used as a gold standard in text simplification for people with disabilities.

4 Assessing the Suitability of Easy-Read Documents for Different Target Populations

Knowing that the majority of the easy-read documents available on the Web comply with the rules of easy-read production guidelines is not enough to accept them as apt for all types of readers with disabilities without evaluating their suitability: a text which is too simplistic or too challenging for a particular group of readers may cause them to lose interest in the text and may diminish their motivation. We evaluate the suitability of easy-read documents with respect to readers with autism (Section 4.1) and readers with mild ID (Section 4.2) by comparing them with corpora evaluated by these readers (LocalNews corpus in the case of ID) or developed by experts (FIRST corpus in the case of autism). If the easy-read sample is significantly more complex or simplistic than the texts in the FIRST and LocalNews corpora, its suitability as a gold standard for accessible texts for readers with autism might be disputed based on the level of simplification the users (LocalNews) and the experts (FIRST) have perceived as suitable for the relevant populations.

4.1 Autism

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised with impairment in social interaction and communication, which influences the language comprehension abilities of the affected individuals (APA, 2013). In previous research we evaluated the suitability of 7 randomly selected easy-read documents from the same 150 document sample used in this study on 20 adults with autism (without intellectual disability) and 20 non-autistic adults matched for age and level of education (Yaneva, Temnikova and Mitkov, 2015). The level of comprehension was assessed through: 1) multiple choice questions, 2) reading times and 3) self-reported difficulty measures for each document. The results indicate that all documents were well understood by all participants, with the autistic participants requiring significantly more time to read them compared with the non-autistic ones. In addition, the autistic participants rank the texts from predominantly “very easy” and “easy” to “moderate” and in a few cases “difficult”, while the vast majority of the non-autistic participants rank them as “very easy”. The study concluded that easy-read documents are understandable enough for adult readers with autism (without intellectual disability), while their perceived level of difficulty is not so trivial as to bore the readers.

As mentioned in Section 2, the FIRST corpus consists of 25 original texts and their parallel simplified versions from the genres of news, education and popular culture. It has been produced by experts specifically working with autistic adults but has never been evaluated by its target population. Our easy-read sample, on the other hand, has been partially evaluated with participants (7 documents only (Yaneva et al., 2015))

	Sent length	Word length	Freq	AOA	Famil-iarity	Concr	Imaga bility	Pas-sive	Nega-tion	1st pers.	2nd pers.	Flesch	Flesch -Kinc
FIRST vs.ER	**-4.4**	-0.1	-1.4	-1.5	**-3**	-0.01	-0.01	-1.8	-0.2	-0.6	**-3.6**	-1.7	**-3.5**
LN vs. ER	**-2.9**	-0.5	**-2.7**	-0.5	-1.7	-1.3	-1.3	-1.8	-0.2	-1.6	-1	-0.6	-1.8

Table 2: Wilcoxon test Z scores for the FIRST and LocalNews (LN) corpora compared to the Easy-Read (ER) sample

and has been produced for people with disabilities as a primary purpose. The comparison of the two corpora is based on the same 13 features as described in Section 3.2.

A Shapiro-Wilk test showed that the data was non-normally distributed, so a Wilcoxon paired signed rank test was applied to compare the differences between the two corpora. Table 2 shows the results of the Wilcoxon test, where the z scores marked in bold indicate 0.001 level of significance.

The main difference between the FIRST corpus and the easy-read texts is that the sentences in FIRST are significantly longer than the ones in the easy-read sample. This difference in sentence length is also the reason why the Flesch Reading Ease formula does not find a significant difference between the levels of difficulty of the two corpora, while Flesch-Kincaid distinguishes between their levels of difficulty, due to its subtler sensitivity to sentence length. The lexical component in both corpora is equally simple, except the fact that the words in the easy-read documents have a higher familiarity level. Finally, the FIRST corpus does not contain many instances where the readers are addressed by second person pronouns, but this could be attributed to the lack of instructional texts in the FIRST corpus compared with the easy-read sample.

4.2 Mild Intellectual Disability

Intellectual Disability (ID) is a condition involving impairment in the general mental abilities of the affected individuals (APA, 2013). The LocalNews parallel corpus (Feng et al., 2009) contains 11 newspaper articles simplified by experts working with adults with mild intellectual disability (MID). Unlike FIRST, LocalNews has been evaluated by 20 adults with MID (Feng et al., 2009). In order to avoid genre bias we only compare the LocalNews corpus against 50 easy-read newspaper articles from our sample. A Shapiro-Wilk test identified the data as non-normally distributed, which is why, similar to the experiment with FIRST, a Wilcoxon signed-rank pair test was applied. The z scores for all 13 features are summarised in Table 2.

Similar to the results from the comparison with the FIRST corpus, again the average sentence length for each document from the easy-read sample is shorter than the average sentence length in the LocalNews corpus, though not to the extent to influence the Flesch-Kincaid formula, which in this case did not differentiate significantly between the two groups of texts. The only other difference is the fact that the words in the easy-read sample had higher frequency scores than those in the LocalNews corpus.

5 Conclusions and Future Work

The results of the presented studies showed that easy-read documents, which were randomly accessed from various domains on the Web, such as charity organisations and government or healthcare websites, comply with the accessibility standard set in the easy-read guidelines. Second, these texts did not exceed the level of difficulty of corpora previously used as a gold standard for accessible writing for autism and mild intellectual disability. Quite the opposite, a presence of shorter sentences and more familiar words was shown, but these did not influence the indices to an extent that would put the easy-read documents in a whole new class of documents, which might be deemed as too simplistic. By satisfying the prerequisites of having good compliance and suitability for autism and MID, easy-read documents show the potential of being a valid gold standard for accessible documents.

Future challenges include exploring the possibility of creating a monolingual comparable corpus of easy-read documents and documents developed for the general audience (e.g. the Conservative Party manifesto versions in Section 1). The creation of such a corpus would allow investigation of ways of aligning parts of these documents where possible, for the purposes of improving automatic text simplification for people with disabilities.

References

David Allen, 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):585–599

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, D.C: American Psychiatric Association.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical methods in Natural Language Processing (EMNLP)*, pages 25–32. ACL.

Conservative Party Manifesto Easy-Read Version. 2015[online] Available at: https://www.mencap.org.uk/sites/default/files/documents/Easy%20Read%20Manifesto_Conservative_V2.pdf. Last accessed: 23.06.2015

Conservative Party Manifesto.2015[online] Available at: https://www.conservatives.com/Manifesto. Last accessed: 23.06.2015

William Coster and David Kauchak. 2011. Simple EnglishWikipedia: A New Text Simplification Task. Proceedings of the *49th Annual Meeting of the Association for Computational Linguistics*: shortpapers, pages 665–669, Portland, Oregon, June 19-24, 2011

William DuBay. 2004. The principles of readability. Costa Mesa, CA: Impact Information.

Equality Act 2010. UK. [online] Avaliable at: http://www.legislation.gov.uk/ukpga/2010/15/section/6. [Last accessed 26.06.2015]

Lijun Feng. 2009. Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS accessibility and computing*, 93, pp. 84-91

Rudolf Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3), 221-233.

Geert Freyhoff, Gerhard Hess, Linda Kerr, Bror Tronbacke and Kathy Van Der Veken (1998) Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability. ILSMH European Association

Ken J. Gilhooly and Robert H. Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*. Volume 12 (4), pp: 395-427

Vesna Jordanova, Richard Evans and Arlinda Cerga-Pashoja. 2013. FIRST Deliverable – Benchmark report (result of piloting task). Technical Report D7.2, Central and Northwest London NHS FoundationTrust, London, UK.

J. Peter Kincaid, Robert. P. Fishburne, Richard. L. Rogers and Brad. S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. CNTECHTRA Research Branch Report 8-75.

Juan Martos, Sandra Freire, Ana Gonz´alez, David Gil, Richard Evans, Vesna Jordanova, Arlinda Cerga, Antoneta Shishkova and Constantin Orasan. 2013. FIRST Deliverable - User preferences: Updated. Technical Report D2.2, Deletrea, Madrid, Spain.

Aurelien Max. 2006. Writing for Language-impaired Readers. CICLing'06 Proceedings of the *7th international conference on Computational Linguistics and Intelligent Text Processing.* pp. 567 – 570.

Danielle S. McNamara, Max M. Louwerse, Z. Cai and Art Graesser. 2013. Coh-Metrix version 3.0. Retrieved [17.05.2015], from http://cohmetrix.com

Misako Nomura, Gyda Skat Nielsen, Bror Tronbacke and International Federation of Library Associations and Institutions (2010) Guidelines for Easy-to-read Materials/rev. IFLA Headquarters, The Hague.

Sarah E. Petersen, and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In Proc. of *Workshop on Speech and Language Technology for Education*, pages 69–72

Sanja Štajner, Richard Evans, Constantin Orasan and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, held in conjunction with LREC 2012. Istanbul, Turkey, May 27, pp. 14-21

World Health Organisation. 2011. World Report on Disability.[online]Available at: http://www.who.int [Last accessed 26.06.2015]

Wei Xu, Chris Callison-Burch and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 283–297.

Victoria Yaneva, Irina Temnikova and Ruslan Mitkov. 2015. Accessible Texts for Autism: An Eye-Tracking Study. To appear in *ASSETS 2015. The 17th International ACM SIGACCESS Conference of Computers and Accessibility*, Lisbon, Portugal, 26-28 October.

Author Index