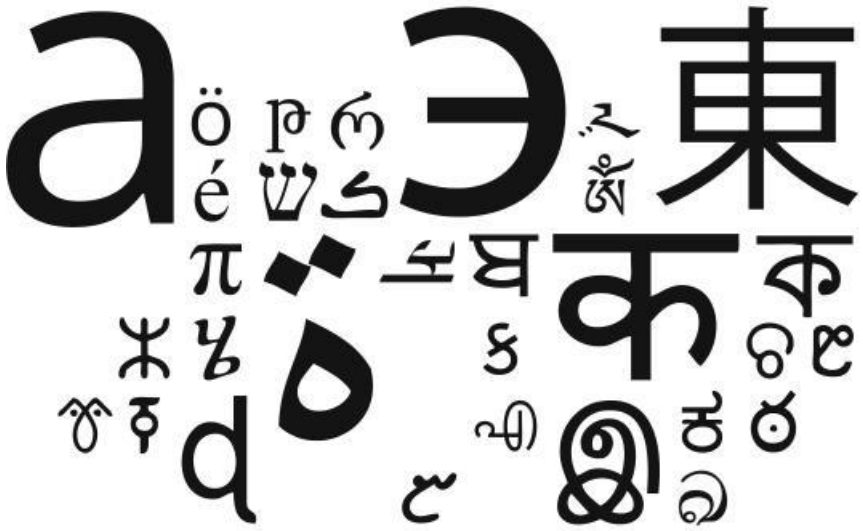


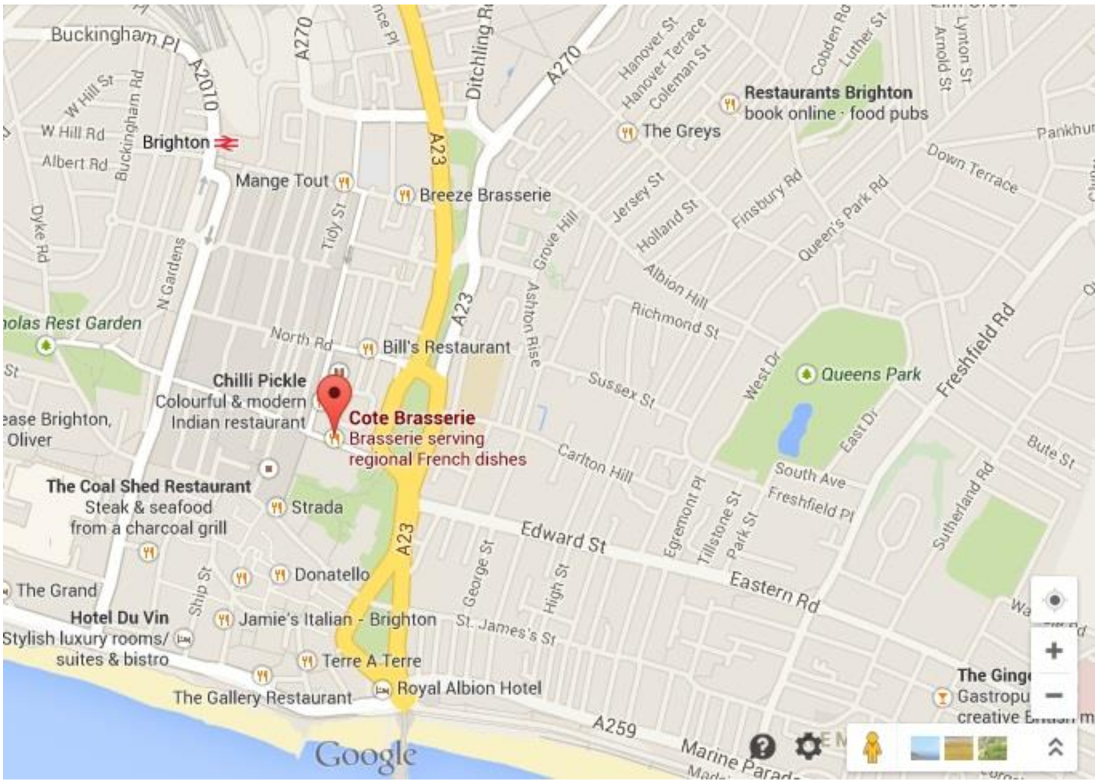
Ninth International Workshop on Written Language and Literacy

September 4th-5th 2014

University of Sussex, Brighton, UK

University of Sussex


East Slope
University Of Sussex
Catering Service
University of Sussex
Great Wood
N - S Rd
Park St
Mill St
Ridge Rd
FALMER
Falmer Hill
Sportcentre Rd
Stony Mere Way
Falmer
East St
American Express
Community Stadium
A27
Lewes Rd
A270
Brighton & Hove
Albion Football Club
Virgin Active - Brighton
Health and Racquets Club
Coldean Wood
The Keep
Stanmer Park
University of Brighton
Park Rd
A270
Lucraft Rd
Google


Buckingham Pl
A2010
A270
Ditchling Rd
A270
Hanover St
Hanover Terrace
Coleman St
Cobden Rd
Luther St
Arnold St
Lynton St
W Hill St
W Hill Rd
Buckingham Rd
Brighton
Albert Rd
Restaurants Brighton
book online · food pubs
The Greys
Mange Tout
A23
Breeze Brasserie
Down Terrace
Dyke Rd
Tidy St
Grove Hill
Jersey St
Holland St
Finsbury Rd
Queen's Park Rd
N Gardens
Albion Hill
Richmond St
North Rd
Bill's Restaurant
A23
Ashton Rise
Chilli Pickle
Colourful & modern
Indian restaurant
Sussex St
West Dr
Queens Park
Freshfield Rd
Cote Brasserie
Brasserie serving
regional French dishes
Carlton Hill
East Dr
Bute St
The Coal Shed Restaurant
Steak & seafood
from a charcoal grill
Strada
South Ave
Freshfield Pl
Edward St
A23
Egremont Pl
Tillstone St
Park St
Sutherland Rd
Eastern Rd
Donatello
George St
High St
The Grand
Hotel Du Vin
Stylish luxury rooms/
suites & bistro
Ship St
Jamie's Italian - Brighton
St. James's St
Terre A Terre
The Gallery Restaurant
Royal Albion Hotel
A259
Marine Parade
Google

Welcome

Welcome to the Ninth International Workshop on Written Language and Literacy hosted by the School of English at the University of Sussex. I am delighted to welcome you to Brighton. This booklet contains the key information about the workshop that you will need as well as the abstracts for the papers to be presented at the workshop.

I hope that you enjoy the workshop and your stay in Brighton.

Lynne Cahill

Locations

The first day will take place in the Jubilee Building on the University of Sussex Campus. The morning and first part of the afternoon will be in room G22, on the ground floor. The afternoon coffee break and poster session will be in room G31, just across from G22.

The second day will take place at The Keep. This is the home of the East Sussex Record Office as well as the Mass Observation Archive of the University of Sussex.

Instructions for getting to The Keep by public transport:

By train

The nearest station is Falmer.

From Brighton – exit platform 2 via the gates, turn left and use the green coloured bridge to cross over the railway. On the other side proceed as if coming from Lewes.

From Lewes – exit platform 1 via the gates, turn right and walk down the wide tarmaced pathway to the very end where there are steps and a ramp. Turn right and go through the short tunnel under the railway and at the end turn left – the entrance to The Keep is in front of you.

By bus

The following Brighton & Hove buses stop just outside The Keep car park at the Brighton Academy bus stop:

- No.23 (Brighton Marina – County Hospital – Queens Park – Lewes Road – Universities)
- No.25 (Palmeira Square – Lewes Road – Universities),
- No.28 (Brighton – Lewes – Ringmer),
- No.5b (Hangleton – Hove – Brighton – Hollingbury – Universities (**peak times only**))

Follow the footpath to your right as you get off the bus into the car park.

The No. 29 (Brighton – Lewes – Isfield – Uckfield – Crowborough – Tunbridge Wells) stops at Falmer train station

The workshop dinner will be at Cote Brasserie in Church St, Brighton. Maps with the locations of The Keep and Cote are on the next page.

On Friday, lunch will be provided, as we will be in a location somewhat remote from places to eat. On Thursday, you are free to find somewhere on campus to eat. The main eating places are in Bramber House, but there are smaller cafés around the campus, including one in the Jubilee Building, where we are based, the Dhaba in Arts C and the Piazza Café, close to Arts A. If you would like a drink over lunch, the IDS (Institute for Development Studies) has a restaurant and bar and Falmer Bar, at the front of campus, serves food as well as drinks.

Programme

Day 1 (Jubilee Building, G22)

9.00 Registration

9.30 Welcome and Introduction

9.45 **Vilma Symanczyk Joppe** "Splitting up German compounds: Writing rules remote from standard orthography"

10.15 **Katarzyna Foremniak** "Punctuation in dictionaries and databases"

10.45 Coffee

11.15 **Dave Roberts** "Building a homograph corpus as a foundation for tone orthography research in Kabiye (Togo)"

11.45 **Kazuhiro Okada** "Typological differences in the Linked Writing of Cursives"

12.15 Oral poster presentations

12.30 Lunch

2.00 **Johan Zuidema and Anneke Neijt** "The BasisSpellingBank: a new description of Dutch orthography based on a triplet lexicon"

2.30 **Lieke Verheijen** "Computer-Mediated Communication: A New Writing System? A Register Analysis of Dutch Written CMC"

3.00 Coffee

3.30 Posters:

Hisashi Masuda, Terry Joyce, Taeko Ogawa, Masahiro Kawakami and Chikako Fujita "A database of transparency ratings for two-kanji Japanese compound words"

Merijn Beeksma, Mijntje Peters, Johan Zuidema and Anneke Neijt "Triplet analysis: converting words into triplets"

Edward Crook and Lynne Cahill "The Influence of Transcription Mode: a comparison of typed and hand-written apology letters"

Mijntje Peters, Johan Zuidema, Anna Bosman & Anneke Neijt "Verb spelling in grade 6: checking, smurfing or just 'practice makes perfect'?"

Coffee and posters will be in Jubilee G31

5.00 AWLL business meeting

7.30 Dinner at Cote Brasserie

Day 2 (at The Keep)

9.30 Invited talk: **Viorica Marian** "An Introduction to On-line Databases Using CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities"

11.00 Coffee

11.30 **Martin Evertz** "Minimal graphemic words in English and German – Lexical evidence for a theory of graphematic feet"

12.00 **Lynne Cahill** "CELEX and PolyOrth: database to lexicons"

12.30 Lunch (optional tour of the Keep and opportunity to view some of the holdings in the mass Observation Archive)

2.00 **Terry Joyce, Bor Hodošček and Hisashi Masuda** "Constructing a database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system"

2.30 Panel discussion on orthographic databases

3.30 Conclusion

Abstracts

Vilma Symanczyk Joppe “Splitting up German compounds: Writing rules remote from standard orthography”

Katarzyna Foremniak “Punctuation in dictionaries and databases”

Dave Roberts “Building a homograph corpus as a foundation for tone orthography research in Kabiye (Togo)”

Kazuhiro Okada “Typological differences in the Linked Writing of Cursives”

Johan Zuidema and Anneke Neijt “The BasisSpellingBank: a new description of Dutch orthography based on a triplet lexicon”

Lieke Verheijen “Computer-Mediated Communication: A New Writing System? A Register Analysis of Dutch Written CMC”

Martin Evertz “Minimal graphemic words in English and German – Lexical evidence for a theory of graphematic feet”

Lynne Cahill “CELEX and PolyOrth: database to lexicons”

Terry Joyce, Bor Hodošček and Hisashi Masuda “Constructing a database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system”

Hisashi Masuda, Terry Joyce, Taeko Ogawa, Masahiro Kawakami and Chikako Fujita “A database of transparency ratings for two-kanji Japanese compound words”

Merijn Beeksma, Mijntje Peters, Johan Zuidema and Anneke Neijt “Triplet analysis: converting words into triplets”

Edward Crook and Lynne Cahill “The Influence of Transcription Mode: a comparison of typed and hand-written apology letters”

Mijntje Peters, Johan Zuidema, Anna Bosman & Anneke Neijt “Verb spelling in grade 6: checking, smurfing or just 'practice makes perfect'?”

Splitting up German compounds: writing rules remote from standard orthography

Vilma Symanczyk Joppe

Bergische Universität Wuppertal , Germany

In the orthography of German, writing [expressions] together or separately (WTS) remains a constant problem area. According to standard assumptions, WTS is a mere reflex of word status: If the expression in question is, morphologically and syntactically, classified as a single word, it must not be broken up by spaces; if it consists of more than one word, spaces must be inserted between them. Consequently, doubts about an expression´s WTS should only arise if relevant criteria for word status do not converge.

Compounds, which form a considerable part of the German lexicon, do not belong to these debatable cases. While constructions of unclear word status, e.g. particle verbs, are discussed extensively in both normative and linguistic literature, the WTS of compounds is often not even mentioned. Recently, however, linguists have noticed1 an increase of “separated compounds”– some kind of spelling variation, which has emerged in violation of unambiguous orthographic rules.

The first part of my talk shall give a brief sketch of the facts outlined above. In the second part, I shall present an own empirical study2 of the phenomenon of separated compounds, based on a corpus of approximately 15,000 compounds from texts with low norm orientation (mainly internet and advertising texts). I will show that compound-internal spaces (1) are quiet common, at least in certain genres, (2) can be triggered systematically by several factors (to the degree that they may even represent the unmarked case), (3) are part of a hierarchy of word-internal structure indicators, and (4) are even produced by orthographically proficient writers, if certain relevant factors are present. I shall conclude that a picture of German WTS which relies exclusively on data drawn from spelling books, normative grammars and texts with high norm orientation would remain incomplete.

1 See, for instance, C. Dürscheid (2000): Verschriftungstendenzen jenseits der Rechtschreibreform. *Zeitschrift für Germanistische Linguistik 28/2,* 237-247; also J. Jacobs (2005): *Spatien. Zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch.* Berlin: de Gruyter (cf. pp. 6; 168).

2 An expansion of the study I did in my M. A. thesis, V. Symanczyk Joppe (2011): *Spatien innerhalb deutscher Komposita. Eine empirische Untersuchung normwidriger Schreibungen,* BUW.

Punctuation in dictionaries and databases

Katarzyna Foremniak

University of Warsaw, Poland

The rules of punctuation, both rhetorical and the syntactical, have been traditionally presented as a description of contexts assigned to individual punctuation marks (Angełowa 1985; Parkes 1993; Mortara Garavelli 2008). As such they have been included to grammar books, rhetorical treatises, syntax handbooks, stylistics handbooks and language usage guides. However, one of the most typical sources of Polish punctuation norm are the punctuation dictionaries, which delineate the rules according to the use of certain expressions, conjunctions, pronouns and syntactic structures. In the presentation I would like to discuss the most popular Polish punctuation dictionaries (see *Table 1, overleaf*) and examine the following questions.

1. How can be lexical-databases and corpus resources used in the preparation of such dictionaries?

2. What kind of data about the use of punctuation marks can we derive from language corpora?

3. How and to what extent can the corpus data be employed to create general rules and principles?

4. Is it possible to describe every punctuation system in a dictionary (a comparative analysis of grammatical Polish punctuation and rhetorical Italian punctuation)?

References

Angełowa, Iskra. 1985. *Charakterystyka interpunkcji polskiej w świetle normy i praktyki*. Zakład Narodowy im. Ossolińskich.

Mortara Garavelli, Bice. 2008. *Storia della punteggiatura in Europa*. Laterza.

Parkes, Malcolm Beckwith. 1993. *Pause and Effect: An Introduction to the History of Punctuation in the West*. University of California Press.

1.	M. Froelichowa, *Słowniczek interpunkcji i zasady przestankowania*, Wydawnictwo S. Arcta, Warszawa 1951.
2.	E. i F. Przyłubscy, *Gdzie postawić przecinek. Poradnik przestankowania ze słowniczkiem*, Wiedza Powszechna 1967. Dziesięć wydań, ostatnie w 1993 r.

3.	J. Podracki, *Słownik interpunkcyjny języka polskiego z zasadami przestankowania*, Wydawnictwo Naukowe PWN, Warszawa 1993.
3a	J. Podracki, *Słownik interpunkcyjny języka polskiego z zasadami przestankowania*, wyd. rozszerzone i zmienione, Wydawnictwo Naukowe PWN, Warszawa 1998 i 1999.
3b	J. Podracki, *Słowniczek interpunkcyjny dla najmłodszych*, PWN Wydawnictwo Szkolne, Warszawa 1999.
3c	A. Gałązka, J. Podracki, *Kieszonkowy słowniczek interpunkcyjny PWN*, Wydawnictwo Naukowe PWN, Warszawa 2001.
3d	J. Podracki, *Słownik interpunkcyjny PWN (mały)*, Wydawnictwo Naukowe PWN, Warszawa 2002.
3e	J. Podracki, *Nowy słownik interpunkcyjny języka polskiego z zasadami przestankowania*, Świat Książki, Warszawa 2005.
3f	Podracki J., Gałązka A., *Gdzie postawić przecinek. Poradnik ze słownikiem*, Wydawnictwo Naukowe PWN, Warszawa 2010.
4.	A. Dzigański, *Praktyczny słownik interpunkcyjny*, Zielona Sowa, Kraków 2004.
4a	A. Dzigański, *Podręczny słownik interpunkcyjny*, Zielona Sowa, Kraków 2005.
4b	A. Dzigański, *Słownik interpunkcyjny*, Zielona Sowa, Kraków 2005, 2007 i 2008.
4c	A. Dzigański, *Nowy słownik interpunkcyjny*, Krakowskie Wydawnictwo Naukowe, Kraków 2009.

Table 1. Polish punctuation dictionaries

Building a homograph corpus as a foundation for tone orthography research in Kabiye (Togo)

David Roberts

The standard orthography of Kabiye, a Gur language of Togo, does not mark tone. In such a context, how can a researcher adequately assess the degree of ambiguity in the written language and make a valid contribution to the debate about how tone might be incorporated in the second generation of language development? This paper approaches this question, not from the perspective of phonological analysis that has tended to dominate the literature, but from the point of view of the linguistics of writing. It advocates the development of a computerized homograph corpus as the first of three analytical steps, and explains how such a corpus can be generated. Adapting Catach's model of lexical ambiguity for Kabiye, three criteria are established – semantics, morphology and dialect variation – on the basis of which the various words and affixes are included or excluded. The homograph corpus can then serve as a basis for the second and third steps: an analysis of ambiguity in natural written contexts, and an error analysis of oral reading performance.

Typological Differences in the Linked Writing of Cursives

Kazuhiro Okada

Hokkaido University, Japan

Pre-modern writings often developed a cursive style, and further, linked writing, in which graphemes are linked to one another. Cursive styles with linked writing triggered the development of numerous scripts, such as Western miniscule alphabets, Arabic cursive Naskh and Japanese hiragana moraic. I will demonstrate that such cursive styles with linked writing differ in terms of the degree of linkage between individual graphemes, and that investigation of this difference requires the use of databases with proper notation. Linkage degree relates to the nature of cursives, and can be further illustrated in terms of 'graphic cohesion' and 'motivation of allographs'. Graphic cohesion is the degree to which graphemes are linked. In *kana-bun*, an old writing style of Japanese, hiragana was presented in a heavily linked manner, but breaks between graphemes did not always represent a split between morphemes. On the other hand, cursive in the Latin alphabet represents the word boundaries with a break of linked strokes. Thus, the graphic cohesion of *kana-bun* is weak, while that of Latin cursive is strong. In terms of motivation for allographs, the cursive of the Latin alphabet developed extensive use of ligatures, i.e., combinations of allographs, which resulted in some ligatures gaining orthographic status: Æ, umlaut, W, to name a few. Whereas *kana-bun* also had many allographs, it has few ligatures and the motivation for allographs lies elsewhere. As a whole, the linkage degree of *kana-bun* is weak while that of the Latin alphabet cursives is strong. Naskh lies between the two, in that it has plenty of conditional allographs, but linkage and breaking do not necessarily relate to word boundaries. Deeper and broader research on this issue naturally requires that such linkage is encoded in orthographic databases, noting in what manner they are linked, e.g., merged or not.

The BasisSpellingBank: a new description of Dutch orthography based on a triplet lexicon

Johan Zuidema & Anneke Neijt Radboud University Nijmegen, The Netherlands

Phoneme-to-grapheme conversion in Dutch orthography has been described in terms of derivational rules inspired by generative phonology (Nunn 1998). The BasisSpellingBank (BSB) follows an alternative approach based on triplets which specify a phoneme or string of phonemes, their corresponding graphemes and the way these two are related, default or not, with or without morphological structure, gemination of letters etc. (Zuidema & Neijt 2012):

(1)

	riet 'reet'		
phonemes	[r	i	t]
graphemes	r	ie	t
relation	dv	dv	dv

	dimensie 'dimension'						
d	i=	[m	E	n]=	s	i	
d	i=	m	e	n=	s	ie	dv
dv	3.3	dv	dv	dv	3.13	4.4	=

[…]	main stress
dv	default value
=	syll.boundary

The first triplet in (1) is {[r,r,dv}, representing {phoneme, grapheme, relation}. Relationships other than the default value are indicated by a numerical reference to spelling categories which are relevant for spelling instruction (Cranshoff & Zuidema 2005:12-18). All phonemes of *riet* are spelled according to the default, but some phonemes of *dimensie* are spelled differently. In the lowest tier, 3.3 refers to /i/ in non-native words (not spelled with default *ie*), 3.13 refers to /s/ in non-native words (often pronounced as /z/), and 4.4 is for morpheme-final /i/. Roughly, the lower its reference number, the earlier a relationship is learned. Triplets are the smallest strings of phonemes and graphemes relevant for the relations one needs to learn.

The BSB contains triplet descriptions of 100.000 words (18.000 lemmas plus inflected forms) learned at primary school. These words are described in 600.000 triplet tokens, 6000 triplet types, of which only 1000 are frequently used. In our presentation we aim to show the advantages of a triplet analysis. For instance, detailed analysis of spelling tests, cf. (2), the results of test item *gedeeltelijk* 'partly' written by 10 year old children.

ge	
ge	554
gee	2
_	4
1,07%	

d	
d	555
b	1
_	4
0,89%	

ee	
ee	522
i	13
e	19
_	6
6,80%	

l	
l	547
_	12
r	1
2,33%	

te	
te	546
dte	1
t	2
_	4
de	4
ter	2
ten	1
2,50%	

lijk	
lijk	546
lik	2
_	5
lek	3
luk	1
lec	1
lijkijk	1
ijk	1
2,50%	

Cranshoff, B. & Zuidema, J. (2005) *Basisspellinggids.* Utrecht & Antwerpen: Van Dale, and Tilburg: Uitgeverij Zwijsen.

Nunn, A.M. (1998). *Dutch orthography. A Systematic Investigation of the Spelling of Dutch Words.* PhD thesis Nijmegen. LOT dissertation series 6. Den Haag: Holland Academic Graphics.

Zuidema, J. & A. Neijt (2012) *Verkennend onderzoek naar de wenselijkheid en de haalbaarheid van een verrijking van de Woordenlijst Nederlandse Taal ten behoeve van spellingonderwijs.* Rapport in opdracht van de Nederlandse Taalunie.
http://taalunieversum.org/sites/tuv/files/downloads/rapport%20VWS%2015022013.pdf

Computer-Mediated Communication: A New Writing System? – A Register Analysis of Dutch Written CMC

Lieke Verheijen

Radboud University Nijmegen, The Netherlands

In recent decades, computer-mediated communication (CMC) has grown explosively as a means of communication. Because the language of CMC can deviate from standard language conventions (orthographic and grammatical norms are loosened), concerns have been expressed that CMC may degrade youngsters' reading, writing, or spelling skills. But before studying the possible impact of CMC on traditional literacy, one has to establish in what ways CMC language is different and unique. Therefore, I will discuss the findings of a corpus study examining the register of Dutch written CMC, revealing the differences between the informal 'CMC language' of the Dutch youth and their more formal 'school language'. My register analysis includes linguistic features belonging to three dimensions of writing: orthography/spelling ('textisms', i.e. "neographical transformations" from conventionally spelled words (Anis 2007:88)), grammar/syntax (in terms of reductions and complexity), and lexicon/vocabulary (e.g. English borrowings, type-token ratio). Because the most distinctive features of CMC language concern spelling (Crystal 2006), I have made a comprehensive typology of textisms with Dutch examples from the SoNaR corpus ('STEVIN Nederlandstalig Referentiecorpus', Oostdijk et al. 2013). Textisms abound in CMC, because there communicating effectively requires speed rather than correctness (Silva 2011:152). A diverse range of CMC modes are investigated: text messages, microblogs (tweets), chats, discussion lists, IM messages (*WhatsApp*), and social networking sites (*Facebook*). This yields linguistic profiles characterizing the language of different CMC modes, thus confirming that "CMC cannot be treated as one single mode of communication" (Hård af Segerstad 2002:234). The extent to which CMC users deviate from standard language and the degree to which they use particular textisms depends on various factors, including user characteristics. Therefore, the influence of youngsters' age on the divergence of linguistic characteristics used in CMC writings and writings at school is also explored, by distinguishing between CMC by adolescents versus by young adults.

References

Anis, J. (2007). Neography: Unconventional spelling in French SMS text messages. In B. Danet & S.C. Herring (Eds.), *The Multilingual Internet: Language, Culture, and Communication Online* (pp. 87–115). New York, NY: Oxford UP.

Crystal, D. (2006). *Language and the Internet*, second edition. Cambridge: Cambridge UP.

Hård af Segerstad, Y. (2002). *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication*. PhD thesis, University of Gothenburg.

Oostdijk, N., M. Reynaert, V. Hoste, & I. Schuurman (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential Speech and Language Technology for Dutch: Results by the STEVIN Programme* (pp. 219–247). Heidelberg: Springer.

Silva, C. (2011). Writing in Portuguese chats :). A new wrtng systm? *Written Language & Literacy*, *14*(1), 143–156.

Minimal graphematic words in English and German – Lexical evidence for a theory of graphematic feet

Martin Evertz

University of Cologne, Germany

It has been frequently noted in the literature that content words need to consist of at least three letters. This observation is commonly dubbed "three letter rule" and explains the existence of pairs such as *be – bee, in – inn* and the occurrence of silent letters in words such as *woe, owe* and *ebb.* The total number of letters of a word, however, does not seem to be sufficient to capture this regularity. A survey of the CELEX database (Baayen et al. 1995) shows that there are (nearly) no graphematic monosyllabic words consisting of three or more letters that end in a vowel letter (exceptions include words ending in <y> such as *sky*) although there are a reasonable number of phonological monosyllabic words ending in a phonological vowel; however, there are numerous monosyllabic graphematic words in English and German that correspond to the same type of phonological words but end in a digraph, e.g. *Schnee* 'snow', *sea*, have a graphematically closed syllable, e.g. *Kuh* 'cow'*, cow*, or even are graphematically bisyllabic, e.g. *aye, awe, owe*.

I propose that findings like these can be accounted for in a supra-segmental theory of graphematics (Evertz & Primus, 2013); to be precise by a theory of graphematic feet and graphematic weight. Drawing on parallels with the phonological hierarchy and the strict layer hypothesis (Selkirk 1984), I propose that a graphematic word consists of at least one graphematic foot which in turn consists of at least one heavy graphematic syllable. Graphematic weight is determined by the structure of the graphematic syllable. The findings of the survey of the CELEX database are thus in line with earlier experimental findings concerning graphematic weight and the graphematic foot (Röttger et al. 2012, Evertz & Primus 2013).

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX Lexical Database. Release 2 [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Evertz, Martin and Primus, Beatrice (2013): The Graphematic Foot in English and German. Writing Systems Research 5.1, 1-23.

Röttger, Timo B. /Domahs, Ulrike/ Grande, Marion/ Domahs, Frank (2012): Structural factors affecting the assignment of word-stress in German. Journal of Germanic Linguistics, 24, 53–94.

Selkirk, Elisabeth O. (1984): Phonology and Syntax: The Relation Between Sound and Structure. Cambridge.

CELEX and PolyOrth: database to lexicons

Lynne Cahill

University of Sussex, UK

The CELEX lexical database was developed in the 1990s, providing a database of the syntactic, morphological, phonological and orthographic forms of around 120,000 words of Dutch and 50,000 each of English and German. The databases were organised so that lemmas and word forms were stored separately but linked by reference numbers so that, for example, all of the word forms for a particular lemma could be identified and, by cross referencing the different sections, all of the different levels of information could be collated.

This database was used as the basis for the development of the PolyLex lexicons, which included syntactic, morphological and phonological information for around 3000 words of Dutch, English and German. Orthographic information was subsequently added in the PolyOrth project. The PolyOrth project was based on the assumption that the underlying, lexical phonological forms could be used to derive the surface orthographic forms by means of a combination of phoneme-grapheme mappings and sets of autonomous spelling rules for each language. This work is reported in Cahill et al (2013). One of the complications we encountered during the project was the fact that the phonological forms that we used from CELEX were not always genuinely underlying forms so, for example, many unstressed vowels in English were given in CELEX as schwa, which made deriving the orthographic form tricky.

In this paper we first explain the process for deriving (semi-automatically) the phonological lexicons from the CELEX databases. We then discuss the different issues we encountered for each of the three languages in developing the orthographic lexicons and how the CELEX database could have been constructed to make this process easier. Finally we discuss what this process tells us about the relationship between underlying phonology, surface phonology and orthography in these three languages.

Reference

Cahill, Lynne, Carole Tiberius and Jon Herring. (2013) 'PolyOrth: Orthography, phonology and morphology in inheritance lexicons' in *Journal of Written Language and Literacy*, Vol. 16 Issue 2, p146

Constructing an ontology and database of Japanese lexical properties: Handling the orthographic complexity of the Japanese writing system

Terry Joyce[1] Bor Hodošček[2] Hisashi Masuda[3]

[1]Tama University, Japan
[2]Meiji University, Japan
[3]Hiroshima Shudo University, Japan

This paper reports on recent developments in ongoing efforts to construct a comprehensive database of the Japanese lexicon. Joyce, Masuda, & Ogawa (2012; in press) provided an initial introduction by outlining database components at the radical, jōyō kanji and word levels. However, taking inspiration from natural language processing (NLP) trends towards merging the formal specifications of ontologies with lexical resources like WordNet and corpora (Huang et al., 2010; Oltramari et al., 2013), recent work has focused on constructing an ontology of Japanese lexical properties (JLP-O). In addition to establishing a guiding framework for efficiently integrating existing lexical resources using NLP techniques, JLP-O will be a valuable tool for evaluating the theoretical and psychological validities of lexical properties.

There are considerable advantages to drawing on existing frameworks like the lemon model (Lexicon Model for Ontologies; http://lemon-model.net/index.php), but because it was developed for alphabet-using languages such as English and German, a fundamental task is to identify the modifications essential for satisfactorily handling the orthographic complexity of the Japanese writing system. One major modification is to expand on lemon's three sub-classes of lexical entry (word, phrase and part) to include a subclass of characters. In large measure motivated by the role of kanji as the core orthographic building blocks for much of the Japanese lexicon, the character sub-class provides a vital complementary perspective on the database. Another major modification necessary for more complete specification of the JLP-O's lexical entry is to expand on the limited set of properties relating to lexical and/or orthographic variation to more adequately represent the considerable range of orthographic variation observed in Japanese (Joyce, Hodošček, & Nishina, 2011). These key modifications greatly enhance the potential for developing sophisticated querying of the database to aid researchers seeking to fully investigate the rich interconnections within Japanese lexical properties.

A database of semantic transparency ratings for two-kanji Japanese compound words

Hisashi Masuda[1], Terry Joyce[2], Taeko Ogawa[3], Masahiro Kawakami[4], Chikako Fujita[5]

[1]Hiroshima Shudo University, Japan
[2]Tama University, Japan
[3]Tokai Gakuin University, Japan
[4]Osaka Shoin Women's University, Japan
[5]Nanzan University, Japan

Given the morphographic nature of Japanese kanji (Joyce, 2011/2013), the dominant principles of Japanese word formation are morphologically motivated. For instance, 漢字 /kanji/ 'Chinese characters' is a combination of 漢 'Han dynasty' and 字 'characters, letter'. However, due to semantic shifts and varying degrees of polysemy, there are also cases where the relationship between the constituents' meanings and the compound meaning is semantically opaque, such as 泥棒 /dorobō/ 'thief'; a combination of 'mud' and 'stick' respectively.

This paper reports on a web-based survey conducted to collect semantic transparency ratings for a sample of approximately 10,000 two-kanji compound words and the database of results. Ratings on 6-point scales concerning the relationship between the meanings of the respective constituents and the compound meaning were collected from 1,710 students.

Conducted analyses reveal a skewed distribution with 94.4% of the compounds having a strong relationship between the constituent meanings and the compound meaning (ratings 4-6), 4.9% having a strong relationship for only one constituent (ratings 4-6 vs. 1-4), and only 0.7% having weak relationships for both constituent meanings (ratings 1-4). At present respondent levels, however, analyses also indicate some individual variations in the ratings, which may reflect both the ambiguous and/or polysemous nature of some constituents, as well as differences in the lexical knowledge of the respondents. Further analyses are also being conducted to explore the correlations between semantic transparency ratings and other lexical properties, such as morphological family sizes, and phonological neighborhoods. As the data is potentially valuable for various researchers, such as psycholinguists investigating the involvement of morphological information in the lexical processing of compounds, it is being made available as a separate database, but, in due course, it will also be integrated within the larger database of Japanese lexical properties under development (Joyce, Horoscek, & Masuda, this workshop).

Triplet analysis: converting words into triplets

Merijn Beeksma, Mijntje Peters, Johan Zuidema and Anneke Neijt

Radboud University Nijmegen, The Netherlands

In order to gain insight in the relation between phonemes and graphemes, Zuidema developed a description of Dutch orthography in terms of triplets (Zuidema & Neijt 2012). A triplet contains three levels that define this relation by using the smallest possible parts of a word - sometimes a single letter, sometimes more letters, because of reciprocal relations between letters. Triplets combine the orthographic level, the phonemic level and the rule-based level. For example, take the Dutch word for triplets, *tripletten*:

{t ; t ; dv} {r ; r ; dv} {i ; i ; 3.3} {p ; p ; dv} {l ; l ; dv}{ett ; Et ; 4.2a} {en ; @n ; 2.2}

The ';' marks the boundaries between the graphemic, phonemic and rule-based levels (the latter contains numbers that refer to orthographic rules or *dv* when the relation follows the default value). The triplets can be helpful to define the orthographic structure of a word and to determine what kinds of words are alike, in terms of patterns or in terms of specific grapheme-phoneme relations. The triplet analysis makes it possible to obtain a lot of information about words that can easily be interpreted and compared. So how does the process of creating triplets work?

This presentation will show how correct and incorrect words of a spelling test are converted into triplets with the aid of tools developed in Microsoft Excel. The complex rules of gemination and degemination in Dutch orthography will be taken as an example. The aim of this analysis is to show whether or not proficiency in (de)gemination predicts proficiency in verb spelling in Dutch.

Reference

J. Zuidema & A. Neijt (2012). *Verkennend onderzoek naar de wenselijkheid en de haalbaarheid van een verrijking van de Woordenlijst Nederlandse Taal ten behoeve van spellingonderwijs.* Online available: http://taalunieversum.org/sites/tuv/files/downloads/rapport%20VWS%2015022013.pdf

The Influence of Transcription Mode: a comparison of typed and hand-written apology letters

Edward Crook* and Lynne Cahill

University of Sussex, UK

*Now at Brandwatch, Brighton

With the rapid decrease in the use of handwritten letters and the increase in personal communication by email, text message and other social media, some mourn the loss of the personal touch. In this research we investigated the impact of typed and hand-written apology letters. Previous research has suggested that the writing process is affected by the mode in a number of ways, including length of texts, speed or production and completeness of sentences (Berninger et al, 2009). In the study reported here we investigated both the effectiveness of apology letters in each mode and the linguistic differences between letters produced with each mode.

In out experiment subjects were asked to construct a letter of apology for having accidentally dented their neighbour's car. The subjects were 30 native English speakers, balanced for age, gender and how often they wrote by hand and by keyboard. Half the group typed their letters and the other half wrote their letters by hand. The handwritten letters were then typed and the group who had hand-written their letters wrote out the letters that had been typed by the other group. This allowed us to measure to what extent any difference could be accounted for purely by perception of the mode and to what extent the mode of composition affected the creative and formal aspects of the letters.

Our two key hypotheses were that people would view handwritten letters, however initially produced, as more effective and that the letters initially written by hand would appear more effective, even when viewed typed. We asked 45 different subjects to rate the letters in terms of their effectiveness by asking them to answer questions about how blameworthy, how sincere, how serious and how effective the person/apology seemed to them. The results indicated that both hypotheses were supported.

Reference

Berninger V., Abbott R., Augsburger A., Garcia N., 2009. 'Comparison of pen and keyboard transcription modes in children with and without learning disabilities', *Learning Disability Quarterly*. 32; Summer 2009, pp.123-141

Verb spelling in grade 6: checking, smurfing or just 'practice makes perfect'?

Mijntje Peters, Johan Zuidema, Anna Bosman & Anneke Neijt

Radboud University Nijmegen, The Netherlands

For years teachers struggle trying to teach their primary school students how to spell Dutch verbs. Several relatively successful recommendations have been made (Van der Velde 1956, Assink 1983, Zuidema 1988), but still only 60% of the students leaving primary school write 80% of the verb forms correct (whereas the official norm is that 75% write 80% correct).

Nowadays an algorithmic version of the theory underlying Dutch verb spelling is used to teach knowledge about the orthography of Dutch verbs, since research showed this is the most successful way to acquire this knowledge. However, this algorithmic way of deciding how to write verbs may be too slow for advanced writers. We investigated whether students in 6th grade benefit from using shortening strategies. The use of a set of abbreviated algorithms, where students have to check whether a verb form meets certain requirements, is compared with the use of a strategy based on using the verb '*smurfen*' as an exemplar.

There are three main problem types to be found when writing Dutch verb forms: d/t-problems , d/dt-problems and single/double-problems. The strategies were used while writing verb forms from these different problem types. From which strategy do the students benefit the most? Are there any differences in results per problem type? Do students with lower grades in spelling benefit more from a certain strategy? Or is the answer to this spelling problem more simple: practice makes perfect?

References

Assink, E. M. H. (1987). Algorithms in spelling instruction: The orthography of Dutch verbs. *British Journal of Psychology, 79*, 228-235.

Assink, E. M. H. (1985). Assessing spelling strategies for the orthography of Dutch verbs. *British Journal of Psychology, 76,* 353-363.

Van der Velde, I. (1956). *De tragedie der werkwoordsvormen, een taalhistorische en taaldidactischestudie*. Academisch proefschrift, Rijksuniversiteit Groningen / Groningen: Wolters Noordhoff.

Zuidema, J. J. (1988). *Efficiënt spellingonderwijs. Een leer- en expertmodel voor het spellen*. Academisch proefschrift, Rijksuniversiteit Utrecht / Amersfoort, Nederland: Acco.