Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science
Polish Academy of Sciences
Warsaw

Journal of Language Modelling

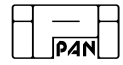
VOLUME 0 ISSUE 1
DECEMBER 2012

Editorials

Articles

JOURNAL OF
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)
ISSN 2299-856X (printed version)
http://jlm.ipipan.waw.pl/



Published by IPI PAN
Instytut Podstaw Informatyki
Polskiej Akademii Nauk
Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Layout designed by Adam Twardoch.
Typeset in XeLaTeX using the typefaces: *Playfair Display*
by Claus Eggers Sørensen, *Charis SIL* by SIL International,
JLM monogram by Łukasz Dziedzic.

EDITORIAL BOARD

Journal of Language Modelling

Adam Przepiórkowski
Institute of Computer Science, Polish Academy of Sciences,
Warsaw, Poland

1 YET ANOTHER JOURNAL?

Welcome to the inaugural issue of the Journal of Language Modelling (JLM), a free open-access peer-reviewed journal aiming to help bridge the gap between theoretical linguistics and natural language processing (NLP).

Setting up a new journal is not a trivial task, and running it possibly for decades requires determination and perseverance, so any such enterprise should not be taken up lightly. The publication of this issue has been preceded by years of growing conviction that there is no appropriate forum for the exchange of ideas between theoretical, formal and computational linguists. Many conversations with our colleagues – both linguists and NLP practitioners – convinced us that such a forum is indeed needed.

Ideally, JLM papers should be accessible to many readers of such periodicals as *Natural Language and Linguistic Theories*, *Journal of Linguistics*, *Language* or *Lingua* on one hand, and *Computational Linguistics*, *Journal of Natural Language Processing*, *Journal of Logic, Language and Information* or *Language Resources and Evaluation*, on the other. The affinity to another relatively young journal, *Linguistic Issues in Language Technology*, should also be clear. On the map of the main linguistic and NLP conferences, we see JLM as close to conferences devoted to constraint-based and formal linguistic theories (HPSG, LFG, TAG, Construction Grammar; Dependency Grammar in general and Meaning-Text Theory in particular; etc.), the *Formal Grammar* conference at ESSLLI, *COLING*, *Treebanks and Linguistic Theories*, etc., but also to *LREC* (*Language Resources and Evaluation Conference*), *TSD* (*Text, Speech and Dialogue*) or the *xTAL* series of conferences (see *Jap-*

TAL 2012, *IceTAL 2010*, *GoTAL 2008*, etc.) on one hand, and *WCCFL* (*West Coast Conference on Formal Linguistics*) or *CSSP* (*Colloque de Syntaxe et Sémantique à Paris*) on the other. The connection between JLM and conferences which concentrate on statistical methods and numerical evaluation, such as *ACL* and *CICLing*, is more complex: results presented at such conferences are relevant to JLM to the extent they say something new about language, not just about usability in an NLP task, and to the extent they are accessible to potentially interested linguists.

It should be clear, thus, that we understand the discipline of language modelling very broadly – much more broadly than normally construed in speech recognition or statistical machine translation. Typical articles published in JLM are expected to deal with linguistic generalisations – their application in natural language processing and their discovery in language corpora – but possible topics range from precise linguistic analyses of phonological, morphological, syntactic, semantic and pragmatic language phenomena to mathematical models of aspects of language, and further to computational systems making non-trivial use of linguistic insights.

2 HOW IT WORKS

JLM is free for everybody – readers and authors alike. All journal content appears on a Creative Commons licence. We strongly recommend the Creative Commons Attribution 3.0 Unported licence (`http://creativecommons.org/licenses/by/3.0/`), which is the default licence for JLM articles,[1] but currently authors have the option to choose a different CC licence.

JLM is published electronically at `http://jlm.ipipan.waw.pl/`, with a printing on demand option (at a fee) coming soon. Once this inaugural issue is published, the journal will receive an ISSN number and every necessary effort will be made to include JLM in all relevant databases and indexes.

Papers are reviewed within less than three months of their receipt (with every effort made to minimise the reviewing period, while

[1] See Stuart Shieber's position paper in this issue on why pure CC-BY should be employed.

maintaining the quality), and they appear as soon as they have been accepted – there are no delays typical of traditional paper journals. Accepted articles are then collected in half-yearly numbers and yearly volumes, with continuous page numbering.

Journal of Language Modelling has a fully traditional view of quality: all papers are carefully refereed by at least three reviewers (including at least one member of the Editorial Board) and they are only accepted if they adhere to the highest scientific, typographic and stylistic standards. It is a contentious issue whether double-blind reviewing is indeed the best reviewing model,[2] but it is suggested by indexing agencies, so we adopt it for the time being. A list of reviewers will be published annually or bi-annually, depending on the turnout (to preserve anonymity).

The current make-up of the Editorial Board (EB) is available at the JLM website. The EB will partially rotate every couple of years, depending mainly on the profile of papers actually submitted to JLM. Day-to-day management of the journal will be performed by selected members of the Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences (`http://zil.ipipan.waw.pl/`), which hosts the journal. Currently, the Managing Editors are Elżbieta Hajnicz, Agnieszka Mykowiecka, Adam Przepiórkowski and Marcin Woliński. The layout of JLM has been designed by Adam Twardoch and the corresponding LaTeX class has been developed by Marcin Woliński, who is also the main Layout Editor. Łukasz Dębowski acts as the Statistical Editor and rotating PhD students act as Copy Editors. All work at JLM is currently performed on voluntary basis and we are counting on community support in order to maintain this business model.[3]

3 THIS ISSUE

This inaugural issue is not a typical JLM issue. It contains two main sections: EDITORIALS and ARTICLES. The former, EDITORIALS, con-

[2] See, e.g., `http://www.linguateca.pt/Diana/SignedReviews.html` for arguments for giving up any anonymity.

[3] Offers from PhD students in Linguistics or Natural Language Processing – English native speakers, to act as Copy Editors, and LaTeX or XeTeX specialists, to help with Layout Editing – are particularly welcome.

sists of three papers: the current editorial, Stuart Shieber's position paper on Creative Commons and Stefan Müller's position paper on Open Access in Linguistics. This way JLM joins the discussion about publishing models in science and makes a clear statement in support of maximal openness.

This issue's ARTICLES section, on the other hand, consists of four papers by partners of CESAR, a European project which sponsors the launch of the Journal of Language Modelling.[4] The first paper, by Radovan Garabík and Mária Šimková, presents the morphosyntactic tagset used in the annotation of the Slovak National Corpus. The next paper, by Svetla Koeva et al., describes design considerations behind the Bulgarian National Corpus. The third paper, by Krešimir Šojat, Matea Srebačić and Marko Tadić, deals with morphosemantic relations between verbs in the Croatian WordNet. The final paper, by György Szaszák and András Beke, is concerned with the the prosody-to-syntax mapping in Hungarian. Note that the languages considered in these papers reflect the linguistic scope of the CESAR project and, similarly, the range of topics reflects the main theme of the project, i.e., language resources. While these articles (and other papers submitted to this issue) underwent the full reviewing procedure, this selection should not be taken as fully representative of JLM, which – as indicated above – has a much broader scope.

4 INSTEAD OF CONCLUSION

Launching a new journal is a high-risk business. Many people have already invested much of their time in this initiative, and returns are far from certain – the community may or may not accept it. We, the Managing Editors, are cautiously optimistic; responses to invitations to join the Editorial Board have overall been very positive, and 9 articles are currently under review for the first regular issue of JLM, to appear in mid-2013. We hope that the often enthusiastic reaction of our colleagues to the idea of JLM is shared by a sufficient number of theoretical, formal and computational linguists to make the current enterprise viable.

[4] CESAR stands for *CEntral and South-east europeAn Resources*; it is a part of the META-NET initiative. See `http://www.cesar-project.net/`.

The Case for the Journal's Use of a CC-BY License

Stuart M. Shieber
School of Engineering and Applied Sciences, Harvard University,
Cambridge, Massachusetts, USA

Scholarly writing is different from other writing. Scholars write articles to disseminate their research results without any thought or desire for direct financial recompense. We write so that science and society can benefit from our insights. We write so that our work can be read, used, and reused.

It bears thinking, then, about how best to make our articles available to the world. Under what conditions should we ideally distribute our articles? The issue is especially apposite in the context of an open-access (OA) journal such as *Journal of Language Modelling* (*JLM*). As an open-access journal, a journal for which online distribution is free, *JLM* is freer to rethink the legal regime under which its articles are distributed, since that distribution does not affect its business model. As a new journal, *JLM* is in the position to design its policies ab initio, without having to worry about past practice or precedent. As a language-related journal, text is both the medium of communication in *JLM* and its object of study, bringing front and center the idea of reuse, especially computational reuse, of the text that comprises its articles.

JLM's staff, in consultation with its editorial board, have thought long and hard about the ideal way to achieve the goal of widest possible use and reuse of its articles. In the scholarly communications community, the consensus view, and the view that *JLM* has settled upon, is to make sure that articles published in the journal are licensed to the world under a broad license that allows every sort of use, subject only to the crucial moral right of proper attribution to the authors. The most direct implementation of that notion is through a Creative Commons license known as CC-BY.

Journal of Language Modelling Vol 0, No 1 (2012), pp. 5–8

When authors provide their work under a CC-BY license, they allow anyone to share their work (copy, distribute, and transmit it), to remix the work (to adapt it in various ways), and to make commercial use of the work. However, any use of the work is subject to an attribution requirement: a user must attribute the work properly to the authors, but may not suggest that the authors endorse their use.

Among the many organizations endorsing CC-BY as the license of choice for OA journals are the Open Access Scholarly Publishers Association, SPARC Europe, SURF, and the Directory of Open Access Journals. The SPARC Europe Seal of Approval for journals even requires CC-BY. All the major OA publishers (Public Library of Science, BioMed Central, Hindawi, and many others) have settled on CC-BY as the license to use, as have essentially all OA experts. Community consensus for CC-BY has been expressed by the authors of the Budapest Open Access Initiative's 10th anniversary recommendation in their crisp statement "We recommend CC-BY for all OA journals." (Budapest Open Access Initiative, 2012) Extended arguments for journals' use of CC-BY have been provided by OASPA (Redhead, 2012) and by Michael Carroll (Carroll, 2011).

Some prospective authors may have concerns about the breadth of the CC-BY license. Such worries are important to assuage.

> *What if someone misuses the material, presenting it in a misleading or inappropriate way, for instance, distributing a version under his or her own name (that is, plagiarizing the work), or providing an inaccurate summary of the work or a bad translation that would reflect badly on me?*

Such uses would violate the CC-BY license. Plagiarism directly violates the attribution requirement of the CC-BY license. Misleading statements or implications that the original author provided or endorses a bad summary or translation similarly violate the license. But more importantly, such misuses violate the social norms of all scholarship, norms that have kept such practices in check throughout the modern history of scholarship. Far more than legalistic remedies, norms of behavior are strong incentives not to misuse others' work. Indeed, if moral suasion is insufficient to stop someone from plagiarism or inap-

propriate attribution, mere legalities of a license are hardly likely to fare better.

> *What if someone starts selling my articles or running other kinds of businesses making use of my writings? Shouldn't I get paid?*

Scholars write for their impact on society, and part of that impact is uptake of their ideas by commercial ventures that improve society through their efforts. Seeing one's work move into the market is a testimony to its importance, not a detriment to be quashed. (As Howard Aiken, the founder of computing research at my own university, has been quoted as saying, "Don't worry about people stealing your ideas. If your ideas are any good, you'll have to ram them down people's throats.")

Keep in mind that although CC-BY allows for commercial reuse, such reuses would need to be something more than simply reselling content. When articles are available for free as in an OA journal like *JLM*, there is essentially no market for pure resale of the articles. Any commercial venture using CC-BY-licensed articles as a part of the business process would need to add value to those raw materials, and insofar as it does so, there would seem to be no argument against legitimate compensation of the business for its efforts in providing that value. If value is added, why not allow recouping of expenses and profit? The knee-jerk reaction against commercial use of scholarly articles has been termed "profit-spite" by Jan Velterop. The sentiment that "if I can't make money off of my article, no one should" may be appealing at first blush, but collapses under an understanding of the scholarly enterprise.

Some of this reaction may be a natural result of popular sentiment against perceived gouging by certain publishers of subscription journals. But the reaction to problems in the subscription journal market is not to blame the publishers, but rather to blame the cause of the systemic market dysfunction, monopolistic ownership. CC-BY eliminates that fundamental problem. When the raw materials for a business are freely available, it's hard for a business to gouge in selling its value-added products and services, because any potential competitor has the same free access to those raw materials.

> *But if someone reuses my article in some way, shouldn't they be required to at least share the results with the community for free?*

Licenses like the "copy-left" license that requires "sharing alike" are appropriate for many situations, especially open-source software projects, where individual modifications of a single item (a software application, say) by themselves can have major value that could otherwise be locked up. But for scholarly articles, any added value would typically come from the ability to aggregate large volumes of articles and extract value from the aggregation. Requiring share-alike would disallow such aggregations, especially when the aggregation includes materials under more restrictive licenses. The overhead of tracking these combined licenses has led many, even in the open-source software community, to eschew share-alike licenses.

I'm proud to be associated with a journal that has made the right choice in ensuring that its articles can be used in the most open and appropriate manner. Journals like *JLM* that act in the best interest of our community of scholars deserve our support.

REFERENCES

Budapest Open Access Initiative (2012), Ten years on from the Budapest Open Access Initiative: setting the default to open, URL `http://www.opensocietyfoundations.org/openaccess/boai-10-recommendations`.

Michael W. Carroll (2011), Why Full Open Access Matters, *PLoS Biology*, 9(11):e1001210, doi:10.1371/journal.pbio.1001210, URL `http://dx.doi.org/10.1371%2Fjournal.pbio.1001210`.

Claire Redhead (2012), Why CC-BY?, Open Access Scholarly Publishers Association Blog, URL `http://oaspa.org/why-cc-by/`.

A Personal Note on Open Access in Linguistics

Stefan Müller
Freie Universität Berlin

ABSTRACT

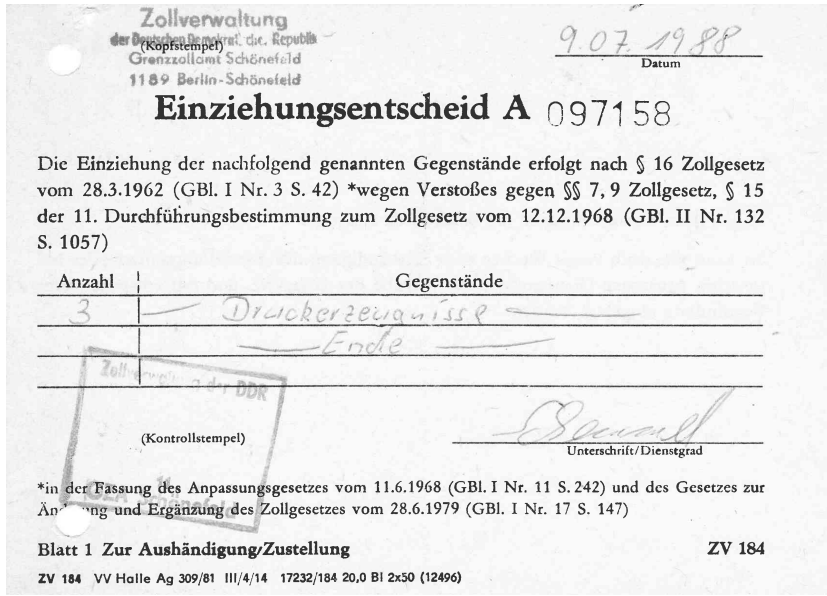
This paper contains only known facts about open access, but they are put into a rather personal perspective that may help others to understand the importance of Open Access in science in general, and in linguistics in particular.

This paper tries to motivate Open Access publishing, with a particular focus on publishing books. In Section 1, I describe the problems in accessing relevant information in economically weak countries, the problem of underpayment in the humanities, and usage restrictions of traditionally published books. Section 2 explains the factors that contribute to book prices. Section 3 briefly describes Open Access publishing and print on demand services. In Section 4, I address some challenges for Open Access publishing and suggest ways to ensure quality control, proper typesetting, and efficient marketing. Section 5 discusses Open Access approaches of profit-orientated publishers. Section 6 deals with Open Access and getting tenure and promotion, and Section 7 is about radical opinions about copyrights outside of academia.

Keywords:
Open Access,
Linguistics,
Precariat,
Rock'n'Roll

1 ACCESSIBILITY OF PUBLICATIONS

In order to make the situation in science more understandable to people living in the US and Western Europe, I want to describe my personal experiences in an Eastern Block country in the following subsection. I will then turn to book prices in relation to underpayment in the



Zollverwaltung
der Deutschen Demokrat. Rep. Republik
(Kopfstempel)
Grenzzollamt Schönefeld
1189 Berlin-Schönefeld

9.07.1988
Datum

Einziehungsentscheid A 097158

Die Einziehung der nachfolgend genannten Gegenstände erfolgt nach § 16 Zollgesetz vom 28.3.1962 (GBl. I Nr. 3 S. 42) *wegen Verstoßes gegen §§ 7, 9 Zollgesetz, § 15 der 11. Durchführungsbestimmung zum Zollgesetz vom 12.12.1968 (GBl. II Nr. 132 S. 1057)

Anzahl	Gegenstände
3	Druckerzeugnisse
	Ende

(Kontrollstempel)

Unterschrift/Dienstgrad

*in der Fassung des Anpassungsgesetzes vom 11.6.1968 (GBl. I Nr. 11 S. 242) und des Gesetzes zur Änderung und Ergänzung des Zollgesetzes vom 28.6.1979 (GBl. I Nr. 17 S. 147)

Blatt 1 **Zur Aushändigung/Zustellung** ZV 184

ZV 184 VV Halle Ag 309/81 III/4/14 17232/184 20,0 Bl 2x50 (12496)

Figure 1: Form confirming that publications were confiscated

humanities even in relatively rich countries in Section 1.2, and turn to usage restrictions for everybody in Section 1.3.

1.1 *Accessibility of Information in Economically Weak Countries or Countries with Censorship*

I was born and raised in the GDR. Since my school had connections to the Humboldt University, I was able to participate in individual tutoring in computer science. This was in 1986 and the subject computer science did not exist back then at the Berlin universities. It was only introduced at the Humboldt University in 1987 as a specialisation of the mathematics curriculum. The individual tutoring was done in the main computing centre of the Humboldt University. The computing centre had one machine with 128 K memory. The lessons were held for a friend of mine and me by two really friendly and knowledgeable guys who did this as their *gesellschaftlich nützliche Tätigkeit.*[1] As a participant in these lectures, I was in a very privileged

[1] People in the GDR were expected to work for society in their spare time. Apart from the implicit pressure to take part in this system, this was a good thing since, among other things, researchers offered classes for pupils in various subjects. The Mathematical Pupils Society still exists: `http://didaktik1.`

situation since I had access to terminals and could type my programs using the editor ED[2]. This was very cool because here I sat as a 16-year-old in front of a screen while the students were typing punched cards,[3] which they then handed in for processing only to find out a week later that they had a typo on the tenth card. However, there was a downside to working at the terminal in the Humboldt University: after typing away for ten minutes, the system crashed and took 20 minutes to reboot. Fortunately, my school had connections to another scientific institution in Berlin: the *Zentralinstitut für Kybernetik und informationsverarbeitende Prozesse* (ZKI, Central Institute for Cybernetics and Information Processing). This institution had a Unix group for which I worked as part of the school education (*Wissenschaftlich-praktische Arbeit* WPA[4]). Here I could work (using ED) on an SM-4[5], a Polish PDP-11 clone.

Now for the publications: of course there were no text books on computer science available at that time in the GDR. Remember, the subject did not even exist back then. So what could people do to get the information they needed? Books? Journals? Magazines? They had to get it via private channels from the West. This was bad luck for me, since I did not have any relatives in West Germany. One could try and bring material in via Hungary, but this was as unpleasant as getting it over from West Germany: if you got caught at the border, the material got confiscated. See Figure 1 on the facing page. In the 50's you even went to prison if you were caught with the wrong type of publication. For example, Manfred Bierwisch was arrested for possessing the journal *Der Monat* and was sentenced to 18 months and actually spent 10 months in prison.[6]

While working at the Humboldt University, I had access to Kernighan and Ritchie, 1978, a book that describes the programming language C. This book was somehow imported from the West and then typed in at the University of Karl-Marx-Stadt (the city with

---

mathematik.hu-berlin.de/index.php?article_id=11.

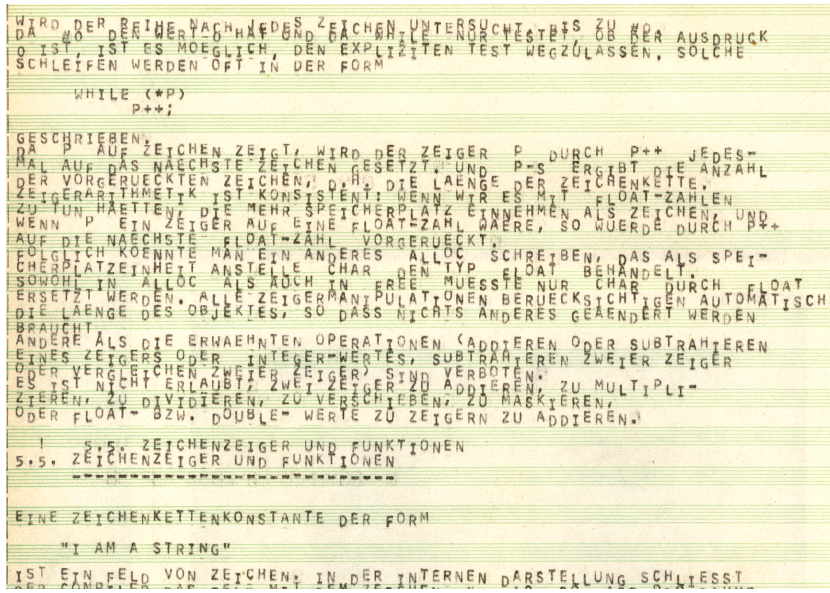
[2] http://www.gnu.org/fun/jokes/ed-msg.html. 19.09.2012.

[3] http://en.wikipedia.org/wiki/Punched_card. 19.09.2012.

[4] http://de.wikipedia.org/wiki/Wissenschaftlich-praktische_Arbeit. 19.09.2012.

[5] http://en.wikipedia.org/wiki/SM-4. 19.09.2012.

[6] http://de.wikipedia.org/wiki/Manfred_Bierwisch. 23.09.2012.


WIRD DER REIHE NACH JEDES ZEICHEN UNTERSUCHT BIS ZU NO
DA #0 DEN WERT 0 HAT UND DAHER NUR TESTET, OB DER AUSDRUCK
O IST, IST ES MOEGLICH, DEN EXPLIZITEN TEST WEGZULASSEN. SOLCHE
SCHLEIFEN WERDEN OFT IN DER FORM

WHILE (*P)
P++;

GESCHRIEBEN.
DA P AUF ZEICHEN ZEIGT, WIRD DER ZEIGER P DURCH P++ JEDES-
MAL AUF DAS NAECHSTE ZEICHEN GESETZT. UND P-S ERGIBT DIE ANZAHL
DER VORGERUECKTEN ZEICHEN, D.H. DIE LAENGE DER ZEICHENKETTE.
ZEIGERARITHMETIK IST KONSISTENT: WENN WIR ES MIT FLOAT-ZAHLEN
ZU TUN HAETTEN, DIE MEHR SPEICHERPLATZ EINNEHMEN ALS ZEICHEN, UND
WENN P EIN ZEIGER AUF EINE FLOAT-ZAHL WAERE, SO WUERDE DURCH P++
AUF DIE NAECHSTE FLOAT-ZAHL VORGERUECKT.
FOLGLICH KOENNTE MAN EIN ANDERES ALLOC SCHREIBEN, DAS ALS SPEI-
CHERPLATZEINHEIT ANSTELLE CHAR DEN TYP FLOAT BEHANDELT.
SOWOHL IN ALLOC ALS AUCH IN FREE MUESSTE NUR CHAR DURCH FLOAT
ERSETZT WERDEN. ALLE ZEIGERMANIPULATIONEN BERUECKSICHTIGEN AUTOMATISCH
DIE LAENGE DES OBJEKTES, SO DASS NICHTS ANDERES GEAENDERT WERDEN
BRAUCHT.
ANDERE ALS DIE ERWAEHNTEN OPERATIONEN (ADDIEREN ODER SUBTRAHIEREN
EINES ZEIGERS ODER INTEGER-WERTES, SUBTRAHIEREN ZWEIER ZEIGER
ODER VERGLEICHEN ZWEIER ZEIGER) SIND VERBOTEN.
ES IST NICHT ERLAUBT, ZWEI ZEIGER ZU ADDIEREN, ZU MULTIPLI-
ZIEREN, ZU DIVIDIEREN, ZU VERSCHIEBEN, ZU MASKIEREN,
ODER FLOAT- BZW. DOUBLE- WERTE ZU ZEIGERN ZU ADDIEREN.

! 5.5. ZEICHENZEIGER UND FUNKTIONEN
5.5. ZEICHENZEIGER UND FUNKTIONEN

EINE ZEICHENKETTENKONSTANTE DER FORM

"I AM A STRING"

IST EIN FELD VON ZEICHEN. IN DER INTERNEN DARSTELLUNG SCHLIESST

Figure 2: Kernighan and Ritchie as it was read in the GDR

three ‘o’s, now called Chemnitz). The sysadmins had a copy of the file and printed it for me on their parallel printer. There was a problem with the printer though: the type wheels were out of sync and so the letters were dancing. See Figure 2. But the good thing was: I read the book in the tram and these trams[7] were shaking so wildly that the dancing of the letters was counterbalanced.

The situation with regard to scientific magazines in general and computer magazines in particular was similar. One could read the Chip (back than the best computer science magazine in Germany) in the *Stadtbibliothek* (Central Library of Berlin) but only as Micro Fiche[8]. My friend wrote a letter to the *Berliner Zeitung* in which he complained about the access restriction for western journals (Höpfner, 1985). He was then invited to talk to the director of the library and the two of us were granted access to the reading room were we could read computer magazines. (Unfortunately, this did not affect other access restrictions, the eight volumes of *Nakayamas Karate Perfekt* stayed out of reach, you had to have a trainer licence to get them). We later discovered that the library in the ZKI also had the journals we were interested in and

[7] The trains must have been of the type TE 59: `http://de.wikipedia.org/wiki/Geschichte_der_Stra%C3%9Fenbahn_in_Berlin#Fahrzeuge`

[8] `http://en.wikipedia.org/wiki/Microform#Media`. 19.09.2012.

we even could take them home over the weekend (and impress girls with reports about SS20 which were published in *Bild der Wissenschaft* although the existence of these missiles was denied in the GDR. Some censor must have failed terribly). Later we hacked ourselves into the library system. Being able to manipulate the system, we could have kept the books, journals, and magazines for ever, but, since we are honest people, we returned everything and sent the librarian greetings on International Women's Day from her own account. I guess the journals were much more useful in the library than they would have been in our flats since, by returning them to the library, apart from being honest, we ensured that a lot of other people could read about the SS20s.

So much for my (pre-)scientific life in the GDR. It will serve as background information for the discussion of recent developments. The following subsections deal with prices of books and publications (Section 1.2), restrictions for usage (Section 1.3), and double payment (Section 1.4).

1.2 *Buying Books and Salaries in the Humanities*

As noted by all scientists, the prices for journals and books are constantly increasing. In April 2012, Harvard encouraged its faculty members to publish in Open Access media only and to resign from editorial boards of media that is not Open Access.[9] Harvard is the richest university in the world but cannot afford the ridiculous amounts of money required to keep their well-stocked library up to date. Imagine what these costs mean for libraries in countries like Poland and Romania. Even institutions in rich countries like Germany cannot afford this since the educational system is notoriously under-financed. I keep getting emails from our library asking which subscriptions can be cancelled. The cuts in education affect positions of technical assistants, research assistants and professors. If we have to choose between people and journal subscriptions, we go for people of course.

But the price problem is not restricted to journal subscriptions, it also affects books. Let's look at some concrete examples. In linguistics, you can find paperback books costing as much as $175/125€

[9] `http://www.guardian.co.uk/science/2012/apr/24/harvard-university-journal-publishers-prices`. 20.09.2012.

(for instance Czepluch, 1996[10], 376 pages, $0.46/0.33€ per page) or hardback books costing as much as $273/195€ (Pasch *et al.*, 2003[11], 816 pages, $0.33/0.24€ per page). The two volumes of the Handbook Syntax by De Gruyter (Jacobs *et al.*, 1993[12], 1029 pages; Jacobs *et al.*, 1995[13], 611 pages) cost $1006/718€ in total ($0.61/0.44€ per page). In Saarbrücken, I could not access this book in the CoLi library because it had been stolen … Another example of a reference work is *The Encyclopedia of Language & Linguistics* published by Elsevier. The second edition has 9,000 pages and costs $6,845/4,151€.[14] This is $0.76/0.46€ per page. You could argue that nobody except libraries would want to buy a 9,000 page book, but if we look at the prices for individual papers from this book, it gets worse: a paper with 8 pages costs $31,50 ($3.94 per page).[15]

Until now I have been using the German translation of an introduction to logic by Allwood, Anderson, and Dahl (Allwood *et al.*, 1973, 112 pages) for teaching logic to linguists and future German teachers. The book was published by Niemeyer and sold for 9,40€ as a paperback. Niemeyer was taken over by De Gruyter and now this book is sold as a reprint for $126.00/89,95€[16] ($1.12/0.80€ per page).

Books are the tools of scientists. We need them. For instance, a Handbuch (handbook) is something you use frequently. Few people will buy a book for over 100€. I buy a lot of books, but my personal limit for a single book is 60€. Undergrads will not be able to buy a 112 page text book for 90€. They just can't. So, who is supposed to buy these books? Let's have a look at the living conditions in the EU. The salaries for academic staff differ a lot from EU state to EU state. While professors and researchers are relatively well paid in Germany, they get half as much in Greece. According to Wikipedia, a German assistant professor (W1) gets a minimum of 3926,84€.[17] A research assistant on pay-grade TLV-13 living in West Germany gets a mini-

[10] `http://dx.doi.org/10.1515/9783110955309`. 19.09.2012.

[11] `http://dx.doi.org/10.1515/9783110201666`. 19.09.2012.

[12] `http://www.degruyter.com/isbn/9783110203417`. 19.09.2012.

[13] `http://www.degruyter.com/isbn/9783110203301`. 19.09.2012.

[14] The web page of Elsevier contains several broken links and no prices, so I took the prices from amazon.com/amazon.de.

[15] `http://dx.doi.org/10.1016/B0-08-044854-2/01999-4`. 24.10.2012.

[16] `http://www.degruyter.com/isbn/978-3-11-096350-2`. 19.09.2012.

[17] `http://de.wikipedia.org/wiki/Besoldungsordnung_W`. 19.09.2012

mum of 3186,61€.[18] While this looks good on paper, the reality in the humanities is different: a lot of researchers have half positions, that is, they get 1593,31€. After taxes and social and health insurance they are left with 1105,02€. If you compare this with the income of a cleaner who gets 10€ per hour, you will be shocked: the cleaner gets 1733€ per month.[19] On the other hand, this is reassuring: you do not have to worry about checking your kids' homework any longer.

Let's think about the things you can do with 1105,02€: the minimum you need to stay alive in Germany is assumed to be 374€ (Hartz 4/Sozialhilfe = state welfare). Depending on where you live you might get an additional amount for rent and heating. In Berlin, which is a really cheap city, this is 394€. People whose flats are more expensive than this are requested to move to cheaper flats (24.700 households had to reduce their living costs and 1.300 actually were forced to move to new flats in Berlin in 2012[20]). So the real minimum is in fact 768€ (= 374 + 394). For comparison, the money that is spent by the 20 % of the German population that have the lowest income is 483€ rather than 374€. If you live on the welfare level, you have 337€ of your 1105,02€ left. If you don't, you probably have nothing left. How many books would you buy per month?

Note also that the German university system differs from the system in the U. S.: while 82 % of the academic staff in the U. S. are professors, with 55 % of all academic staff tenured, only 14 % of the staff are professors in Germany and only 12 % of the academic staff are tenured (Kreckel, 2008). Figure 3 on the following page illustrates.[21] This leaves us with 86 % research assistants and 74 % of academic

[18] `http://oeffentlicher-dienst.info/tv-l/west/`. 19.09.2012.

[19] I made the somewhat idealising assumption that the cleaner works 40 hours per week: 40 × 10 × 52 / 12. The payment of cleaners for private flats varies between 7.50 and 15€ depending on where you are in Germany. After the switch from BAT-II to the TLV system in 2006 the salary does not contain a location-dependent component any longer. So while a cleaner in Hamburg or Munich gets more than a cleaner in a smaller city, all researchers get the same salary despite huge differences in living costs.

[20] `http://www.tagesspiegel.de/berlin/kosten-der-unterkunft-werden-neu-geregelt-hartz-iv-empfaenger-duerfen-teurer-wohnen-/6473264.html`. 19.09.2012.

[21] The figure is extracted from `http://www.gew.de/Binaries/Binary65439/WiKo10_Reinhardt_Kreckel.pdf`. 20.09.2012.

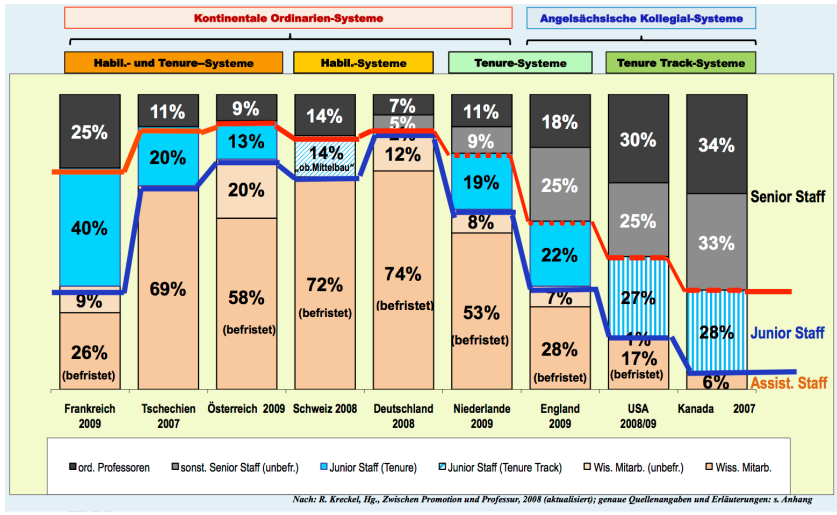


Figure 3: Temporary and permanent positions in different academic systems

staff in non-permanent positions. This means that 74 % of the scientists working at the universities are constantly looking for new jobs, since they have contracts for two or three years. If they fail to get a new job or cannot move to another town or country for whatever reasons, they have to live on social welfare. Again, the cleaner is in a much better position.

It is almost impossible to get permanent positions at German universities below the professor level nowadays. The reason is that the university system is underfinanced in most federal states and universities use the non-permanent positions to save money: whenever one of the temporary positions becomes vacant, it is blocked from being filled for six months to save money. Most German universities constantly run with just 80 % of their academic posts filled. Now the question: would you buy books if you lived under the constant threat of becoming unemployed? If you have 100€ left, you would rather save it for hard times.

Now compare this situation to other countries in the EU. Due to Winfried Lechner's Salary transparency project[22] we know what an assistant professor earns in Greece: 1,361€ after taxes. From this salary you have to subtract at least 150€ per person for private health insurance. What is left is comparable to a research assistant in Germany

[22] `http://users.uoa.gr/~wlechner/`. 19.09.2012.

on a half position. A research assistant in Greece earns 1,291€ after taxes. In comparison to 2008, salaries have been cut by over 26 % (up to 32,2 % for full professors). The living costs in Greece are lower in general, but if you live in Athens, not much is left for buying books.[23]

We have seen some examples from the EU and I leave it to the reader to imagine incomes and living costs of researchers living in other, less wealthy countries.

1.3 *Accessing and Using Books*

The preceding section discussed the impossibility of buying books as an individual scholar, but maybe we do not need to buy books. Maybe Kuczynski (1983) was right in letting his great-grandson ask why he afforded the luxury of a big private library. The alternative is, of course, public libraries. But, again, we have to deal with underfinanced educational and research systems. Some countries have few libraries and if finances are restricted, how would you convince the librarian to buy books about Pirahã? There is no immediately visible economic use of such linguistic research (not even in computational linguistics, since there are just a few speakers of Pirahã and they do not have iPhones).

Apart from this, there is the problem of access restrictions that may be imposed in certain countries. You cannot predict what will be banned for what reasons. I was shocked to find *The Diary of a Young Girl – Anne Frank*, *The Adventures of Huckleberry Finn*, and *Alice's Adventures in Wonderland*, *Harry Potter and the Philosopher's Stone*[24] on the list of banned books maintained at `http://www.banned-books.org.uk/`. I lived in a state where *Animal Farm* (also on this list) was banned: I know of people who got this novel page by page in letters from their relatives in the West. And, as mentioned earlier, in the 50's and 60's you could go to prison for 25 years if you were caught with

[23] According to user information at `http://www.numbeo.com/cost-of-living/compare_cities.jsp?country1=Germany&country2=Greece&city1=Berlin&city2=Athens` the living costs in Athens including rent are 12,36 % higher then the living costs in Berlin.

[24] Banned and burned in some states of the U. S. for promoting witchcraft. See `http://en.wikipedia.org/wiki/Religious_debates_over_the_Harry_Potter_series` for more information on the debate. I guess, depending on the country, *Pippi Longstocking* should also be banned. Either because the main character has red hair or because Pippi is much stronger than the boys.

the wrong books. As should be clear from what I just said: the problem was to get the books. All possible routes into the country were controlled: letters, parcels, the borders. The internet did not exist back then in the Eastern Block countries.[25] But this has changed. People in many countries have ways to connect to the outside. The access to information can take place through ssh tunnels, which will be difficult to decrypt for authorities. The question whether a book may be imported or not will not arise. Whether there will be any conflicts with state authorities depends solely on the diplomacy of those who are accessing the files.

We have seen that there can be economical or political reasons for books being inaccessible, but even for those researchers who have access to books and who can afford to buy them, the situation is unsatisfactory since they cannot work with them properly because buying the book usually does not include the right to access and store an electronic version of the work. Those who need the book in electronic form (for searching or accessing it in places where you cannot take heavy books to) have to scan the book and illegally store the file.

1.4 *Double and Triple Payment*

Writing books, reviewing books for publishers, and copying them afterwards takes a lot of time. This time is usually paid for by state institutions or funding agencies. The publishers do not pay for it. On the contrary, some even require money from the authors to keep the book prices low (depending on the number of copies printed, this can be 1000€).

I want to give two more concrete examples: I spent one month reviewing a book for Benjamins. You may check the sources mentioned above to find out what this actually cost German taxpayers. I am not

[25] We had Unix systems at the Humboldt University and they had a network, but this network was not connected to the outside world. It was nevertheless useful: we had a disk quota of 1 MB. One could not log out of the machine if one used more than 1 MB. The trick was to pack one's files into an archive and send this archive to oneself per email. At the next login the archive was extracted from the mail folder.

It was in the news today (25.09.2012) that the Islamic Republic of Iran will have such an inhouse email system soon. I hope that computer scientists, computational linguists, and others will find ways of using it …

complaining here about the work. I enjoyed it very much. I would have read the book anyway since I found the topic interesting. But the problem is that the publisher now has a product that is better than it would have been without my work. The publisher will sell more copies and will make more money out of it. If I had not asked for it, I would not even have gotten a sample copy of the book. The book costs $150/100€ now and is way too expensive ($0.45/0.30€ per page).

Another example is the books that appear in the legendary HSK series of De Gruyter. Authors do not even get a copy of their book. They get the right to download 10 journal articles (which would cost 400€ if you ever paid for them) and they can buy the book with a 30 % discount. As will be discussed in the next section, the distribution costs for books (distributor, bookshop) make up 55 % of the total cost. Since these costs are saved when authors get their copies directly, De Gruyter takes the normal margin plus an additional margin of 25 % from their authors.

The final example shows that we even have cases of triple payment: Christiane Fellbaum reported a case in which a publisher wanted to charge her (personally) since she wanted to reuse a figure with another publisher. So we have to buy our own products in order to give them to another publisher from which we can then buy it again.

So, we do a lot of work that is paid by the tax payer or not at all and we pay again when we buy our products.

2 WHO DOES WHAT AND WHY IS IT SO EXPENSIVE?

As we saw in the previous section, book prices and journal prices are just too high. But why is this so? Who gets all this money? Depending on the contract with the author, the costs of a book consists of the following:

1. work done by the author (paid by the research institution or by nobody)
2. production costs (editorial process, employees of the publisher, design of cover, ISBN number,[26] paper, printing)

[26] The costs of ISBN numbers vary from country to country. While they are free in Canada, ten ISBN-A cost 100€ and 500 cost 1,500€ in Germany.

3. storage and maintenance cost for infrastructure for electronic publications
4. advertising costs
5. publisher's profit margin
6. royalties for the author
7. storage of the books (Zwischenlager)
8. distribution of the book by a main distributor
9. distribution of the book by a bookshop
10. added value tax (depends on the country/state[27])

Sadly enough, items 7–9 can be up to 55 % of the total book price. As for the author's royalties: we saw in Section 1.4 that authors often pay publishers in order to keep the book price low. When this is the case, they (probably) do not get any royalties. So we are left with the production costs and the margin of the publisher. Here it really depends on the publisher. Some publishers (as, for instance, De Gruyter) do a really excellent job and typeset the whole manuscript so that you end up with really beautifully typeset books. Others, like Niemeyer, didn't do much. They accept submissions in Microsoft Word[28, 29] and I even remember photocopied typewriter manuscripts that had a page 129a between the pages 129 and 130. So the efforts spent on typesetting vary but most publishers do proofreading, check the margins, paragraphs, watch out for widows and orphans, and so on. This involves human labour and hence is expensive.

[27] In Germany the added value tax for food and books is 7 %.

[28] Word is not a typesetting program. It often gets the kerning wrong and there were times when it hyphenated German *aber* as *a-ber* which is correct according to the spelling rules (Duden, 1996, p. 61), but is typographical nonsense.

[29] My favourite story about Microsoft Word is about a paper that was published by Microsoft research (Gamon and Reutter, 1997). I tried to print it (back in 1998) and it caused a paper jam. While preparing this paper, I had a look at the PostScript file again and the PostScript to PDF converter that comes with Mac OS failed on it. Closer inspection of the file revealed that it included printer specific commands.

While editing the HPSG proceedings this year, I almost went nuts since the AVMs in PDF files produced by Microsoft Word caused problems when the PDF was created under Windows. They looked correct under Windows and some non-Windows viewers but were broken on others.

So much for standards and an open world with free communication.

The profit margins are generally unknown but for some large publishers the numbers are available: Elsevier had a turnover of £2,058 M and adjusted operating profit of £768 M in 2011.[30] Springer Science + Business Media S.A. had a turnover of 875 M € and pre-tax profit of 313,3 M €. This is a margin of 37.17 % for Elsevier and of 35.80 % for Springer. This should be compared with Siemens (8.6 %) and Daimler (5.6 %). I clearly remember the outcry that went through the German press when Josef Ackermann, back then the CEO of the German Bank, announced that the German Bank aims for a 25 % of return on equity. Return of equity is different from pre-tax profit and a more realistic goal for a bank, but here we are: our beloved publishers make even ten percent more profit before taxes than what was thought to be *Turbokapitalismus*[31] ('turbo capitalism') in the German Bank discussion.

To conclude this subsection, what we get from good publishers is the following:

1. a brand corresponding to a well-maintained profile
2. marketing (display at conferences, leaflets, web pages, mailings, library contact)
3. support in getting the permissions for reprinting figures (only relevant in some subdisciplines like neurolinguistics)
4. storage and maintenance of electronic publications
5. typesetting
6. proofreading

The publisher also takes care of organisation of content that may affect the market success of a book. We do not necessarily like this as authors …

The costs in 1–6 are fixed costs that do not depend on the number of copies that are printed. If the book production costs 2000€ and you sell 500 copies, the production costs per book are 4€. If you sell 100 copies, the production costs are 20€ per book.

[30] Reed Elsevier Annual Reports and Financial Statements 2011, p. 9. Availible: `http://reporting.reedelsevier.com/staticreports/Reed_AR_2011.pdf`. 24.10.2012

[31] `http://de.wikipedia.org/wiki/Turbokapitalismus`. 24.10.2012.

The result is that publishers are not interested in publishing books with a limited audience (on Danish or formalisation of linguistic theories) or they are interested and charge enormous amounts of money for the printed book.

3 THE SOLUTION

The solution to these problems is Open Access publishing. Publications that are accepted after a selective reviewing process are stored on central storage and archiving servers, for instance those that are provided by university libraries. For books it is still important to have the printed version, so here the solution is Open Access in combination with print on demand services. The copyright is granted by the Creative Commons CC BY[32], which allows the work to be printed and figures to be reused, provided the original work is cited. The translation rights remain with the author. The authors can additionally put a PDF file of their work on their web page and for those of us who prefer printed books, printed copies will be available on demand. In order to get some idea about prices, we can look at the Amazon daughter *Create Space*: as of 20.09.2012 *Create Space* publishes a 15×23cm book ($6'' \times 9''$) with 450 pages for \$10.5 = £8.7 = 10€ (\$0.023/0.022€ per page, cheaper by a factor of 20–50!). This includes a free ISBN number and the number of printed copies is irrelevant for the price calculation. For a price calculator see `https://www.createspace.com/Products/Book/`.

Note that the publication as a printed book is not obligatory. The main publication format is the electronic publication. Those who want to stay away from large companies do not have to print and distribute their book via Amazon. There are several other print on demand services that can be connected to the publication on a web page.

In the next section I address the challenges for and advantages of publishing this way. Some of these challenges have already been addressed by College Publications, which is a publisher run by academics for academics. College Publications was founded by Dov Gabbay and Jane Spurr and has interesting series in Logic, Linguistics, and Computation. They also guarantee low book prices (\$15–25 per an average

[32] `http://creativecommons.org/licenses/by/2.0/`. 27.09.2012.

300-page book) and allow authors to put their book at their web page *once sales have achieved profitability*.[33] As will become clear in the next section, we want to take the whole approach a step further.

4 CHALLENGES AND ADVANTAGES

If we publish without traditional publishing houses, we have the following challenges:

- quality control
  - content
  - proofreading
- typesetting
- marketing
- long-term accessibility of documents

We are currently in the process of setting up Open Access Platform for linguistics books (OALI, `http://hpsg.fu-berlin.de/OALI/`) and I will therefore focus on books in the following subsections.

4.1 *Quality Control*

Publishing houses differ as far as quality control is concerned. Some ask researchers to review book manuscripts, some rely on the reputation of the respective authors and publish whatever they submit. The reviewing process could be organised by the scientific community without the mediation of publishing houses (we do this for conferences anyway). The duty of a reviewer would be to comment on the manuscript and also to point out typos (part of proofreading). Typos could be marked during reading and communicated to the authors via a PDF (either the marked PDF of the submission or a scanned paper version). The names of the reviewers can be published on the book. Publishing the reviewers' names is an old suggestion by Geoff Pullum: Pullum (1984) suggested in his column in NLLT that reviewers of journal articles should be named in the article since this ensures that they take reviewing seriously and also gives credit to their work since more often than not the reviewed piece profits from the comments. We can take Pullum's ideas even further. We can build a web of trust.

[33] `http://www.collegepublications.co.uk/about/`. 24.09.2012.

We can set up reviewing systems that keep the original submission around and add the review. After a revision there could be another review and another revision. Readers can comment and other readers can vote on books and comments. The versioning is basically what we have in Wikipedia and the rating system is practised very successfully on `http://stackexchange.com/`. stackexchange.com has a list of rated questions and answers and you get credits for asking and answering questions. So you can see who is an experienced user of this system. There are certain thresholds for user privileges that are assigned automatically by the system. It may sound silly at first, but it is psychology (Radoff, 2011): doing reviews with such a system is much more fun than doing it just for the love of it. We could give points for reviewers who are fast (you will also have the information about the time the reviewing took, if we keep versions of documents and reviews publicly available). Of course not all reviewers may want this, but we could give points or badges for transparency. So everybody has the option of making her or his review publicly available, but does not have to. There is also the problem of rejected manuscripts. The reviewing work should be credited somehow by the system although the identity of the reviewer does not have to be revealed.

In the case of manuscripts of bad quality, that is, manuscripts that would require a lot of work on the reviewer's side, it could be the case that nobody is willing to review the manuscript. This can be either accepted by the author as a rejection or the author could increase the motivation for reviewing by setting a 'bounty'. Bounties can be set on systems like stackexchange to increase the priority of a question. Those who answer the question will get a bigger number of credit points and the person who asked the question has to 'pay' with some of his or her credit points. If the manuscript has some good ideas in it, reviewers eventually will be willing to invest a lot of time in a manuscript. The extreme version of the 'bounty' idea is of course co-authorship.

Reviewing will normally be done by researchers with a PhD, but the envisioned system allows something like a customer's review, which can be written by everyone. Readers can comment on the books they read and will get points for this. Others can judge the reviews as useful or adequate and this could result in further credits assigned to the author of the review. In that way, talented researchers below PhD level can build a reputation.

Of course, setting up all this in a way that is accepted by the community and that is not vulnerable to manipulation is a non-trivial task. If you are interested in helping to develop and extend software in the directions indicated above, please register at `https://lists.fu-berlin.de/listinfo/OALI-developers`. Issues related to Open Access in Linguistics and the development of the software will be discussed in Frank Richter's blog at `http://www.frank-m-richter.de/freescienceblog/`.

Another interesting aspect in this scenario is that one can use it to bridge the huge gaps in linguistics that some call a crisis of the subject. I think that from a bird's eye view[34] frameworks are not too different (Müller, 2010, Submitted) and maybe a reviewing system that motivates people to look at each other's work critically can bridge the gaps and in the end will result in improved quality in all areas of linguistics.

Since books are printed on demand, there will not be 100 or more copies that have to be sold until one gets to the next edition. This allows for the correction of typos and errors. A version chaos can be prevented by introducing time limits for resubmission.

The big advantage of this publishing model over the traditional one is its flexibility and speed. One could imagine settings in which the author sets the price of a printed book so that it includes royalties for the author. In such a setting we could request that the author pays the reviewer and maybe even an overhead for the organisation. This could be a fixed price to reduce management overheads or a certain percentage of the book price. The authors paying the reviewers seems to be a conflict of interest, but the reviewers are interested in the commercial success of the book and will do everything to improve it and, in addition, their name is published with the book, so they will do whatever they can to ensure quality.

Publishing houses live from their brand names. When they go bankrupt other publishers buy them, just to get established journals and book series (Mouton, Niemeyer, K. G. Saur Verlag → De Gruyter; Kluwer, Springer → Springer Science + Business Media, …). What we need for books is a brand. We are seeking to establish a brand name that is associated with high quality books. The initiative at the FU is

[34] Some birds can fly as high as 11,300 m.

therefore looking for linguists from all branches of linguistics who are willing to participate in the reviewing process and who want to take part in building a brand with a good reputation. The area of expertise of the first supporters of the original proposal is in data-informed theoretical linguistics but, of course, book series in other research areas are possible and welcome.

What we are doing is similar to what publishing houses do. In order to keep the task manageable, we need software for load distribution (the gamification idea mentioned above) and we need subareas of responsibility, that is, series that are run by a few editors that are acknowledged experts in their field. Publishing houses usually have an advisory board that installs such series. We need a board of experts, too. This will be set up in the near future.

4.2 *Typesetting*

While you have full control over your product if you publish without a traditional publisher, the disadvantage of this alternative way is that you have full control over your product. This means that you have to do the proofreading and the typesetting alone.[35] Cambridge University Press estimates the costs for typesetting a 400-page book at £1000. If authors are willing to invest the equivalent of £1000 in formatting the text themselves by learning how to use LaTeX, the nonprofit publisher saves this money and the readers will enjoy cheaper books. One good thing about Open Access publications is that they can be Open Source too. I gave away the source of my HPSG textbook (Müller, 2008), but, while working on the Persian book (Müller *et al.*, In Preparation), I discovered XeLaTeX and also switched to a new tree drawing package. Once this book is finished, I will make the source code available, so that everybody who is interested can learn how certain elements can be typeset. The source code of this article is available at `http://hpsg.fu-berlin.de/~stefan/Pub/oa-jlm.html`.

Frank Richter and Chris Culy are currently working on a translation of LaTeX into e-book formats that would also make it possible to publish typographically complex books including glossed and heav-

[35] Many of us actually prefer doing the typesetting by ourselves instead of having professional typesetters but non-linguists messing around with the symbols. See for instance Ivan Sag's post on the HPSG mailing list: `http://hpsg.stanford.edu/hpsg-l/1997/0062.html`. I typeset all of my books myself.

ily crossreferenced examples, syntactic trees, OT tableaux, and feature structures.

At the OALI kickoff meeting it was decided that Microsoft Word submissions should be an option. So it is like with the *Journal of Language Modelling*: Microsoft Word possible, but LaTeX strongly preferred. Editing and translating Microsoft Word files is labour and hence cost intensive. We have to find ways to pay this work, for instance by external funding from foundations, research organisations, or donors. Of course authors who submit manuscripts that cause a lot of work could also donate or find local people who do the LaTeX conversion for them, which may be a lot cheaper than doing this in countries like Germany.

If we have more donations than are needed for typesetting and other running costs, we could subsidise the printing and even offer books below the print on demand costs.

4.3 *Marketing*

What we cannot do is travel around with a truck of books and present them at conferences. But we can establish a brand and we can promote it at conferences. Actually during show time at the end of our talks. Since the books are accessible, search engines will find them. Since students and researchers from economically weak countries will be able to access them, there will be a bigger audience for the book.

Publishers send books to multipliers, known researchers in the field that may be interested in the book, use it, cite it, and tell others about it. The names of the multipliers are usually provided by the authors. Furthermore, reference copies are sent to certain libraries (for instance the German National Library for books that are published in Germany). In addition, books are sent to journals for reviewing. We will put together lists of journals and help the authors send the books to them.

The most important site for promotion of work in linguistics is the Linguist List. It has over 25.000 readers. Non-profit work can be announced there without any fees.

In addition to the Linguist List publications can be advertised on `www.academia.edu`. As of 06.11.2012, there are 5,704 researchers that have *linguistics* as their research interest and 11,633 with the research interest *languages and linguistics*. When you upload or link papers on academia.edu, you can tag them with respect to certain

research interests. Everybody who has research interests that correspond to the tags will see your new upload. This results in much more targeted distribution of information. For instance, there are 1361 researchers with *Natural Language Processing*, 972 researchers with *Computational Linguistics*, 916 researchers in *Syntax*, and 145 researchers with *Persian Language* among their main research interest. If you tag your paper accordingly, the news about this paper will reach these researchers directly. We have also installed the group *Open Access Books in Linguistics* in which freely accessible books can be announced.

4.4 *Long Term Accessibility of Documents*

The books will be printed and copies will be sent to and archived by reference libraries. Long term accessibility is guaranteed for the electronic publications: (German) university libraries have storage systems that can be used for long term storage. These document servers are connected to international catalogues, which guarantees visibility of the documents. The documents get a Digital Object Identifier (DOI) and, hence, servers going out of business, changing URLs and so on do not cause problems.

4.5 *Advantages*

In the previous subsection we looked at challenges for the new approach and concluded that they are manageable. This brief section highlights some of the advantages. They are:

- Speed
- Version control
- Visibility
- Connection to the primary data and software

The new way of publishing allows scenarios that differ from what we know so far. Authors can make drafts of their book available to the community as soon as there is something to show. Some authors do this and did this in the past, but publishers frowned upon this and some refused to publish books that were in the net before submission. The initial submission could be stored together with reviews and with improved versions of the document (Wikipedia-like). This guarantees that the earliest moment in which a certain idea was present is doc-

umented. It also opens up new possibilities of quality improvement. Readers alert authors of mistakes and typos in a phase after acceptance and before finalisation.

The documents are accessible by all search engines not just by those that are run by companies that happen to have a contract with the publisher.

The publications can be stored together with primary data and software. (Some) publishers are setting up this infrastructure now and some authors have already done this on their own, but there are entirely new possibilities as demonstrated by the projects of Enhanced Digital Publication[36]. The Enhanced Digital Science project looked at various disciplines, also including linguistics. One project developed a dictionary for Berber that includes background information like pictures and articles from the press. Other useful combination of sources can easily be imagined for linguistics. For instance, one could connect a text to corpora and subcorpora that play a role in an analysis, one could include interactive example trees that can be manipulated and modified by the reader since they are connected to online demos of linguistic software. Usually fully worked out linguistic analyses are highly complex. It is the task of the author to simplify the analysis and highlight the most important aspects. However, in some situations the reader wants more or it is not obvious how several partial descriptions in a paper have to be fused into a coherent picture. Having a tree that contains all the details of an analysis but initially displays only the information that was marked as relevant by the author is extremely useful here.

5 OPEN ACCESS AND TRADITIONAL PUBLISHERS

You may wonder why we should go through this organisational nightmare. Aren't there publishers already who publish Open Access, both journals and books? Yes, there are. But, of course, most publishers are profit-orientated companies. Let's look at some examples.

[36] `http://www.surf.nl/en/themas/openonderzoek/verrijktepublicaties/Pages/default.aspx`. 08.11.2012.

5.1 *Elsevier, Springer, De Gruyter*

Elsevier offers the option to publish open access, but charges the authors. Their web page is not transparent, but `http://www.sherpa.ac.uk/romeo/PaidOA.html` lists publishers with an OA option, and according to this page, Elsevier charges $3000 or $5000 for one article, depending on the journal. During a panel discussion on Open Access I asked the Elsevier representative in the panel, Angelika Lex, about the $5000 that Elsevier takes for a publication in *Cell*.[37] I asked about how Elsevier justifies a fee that corresponds to a month's salary of a German professor. The reply was that there are general costs for storage and maintenance and the editorial process and that *Cell* is a very prestigious journal with four articles out of one hundred submissions actually published. So, what Biologists really pay for is the brand. A brand that they helped to establish and that they are helping to maintain by doing the quality control.

Springer offers a publishing model in which all rights remain with the author.[38] However, the authors or their institutions have to pay $3000/2000€ (excl VAT).

Finally, let's look at De Gruyter. At the beginning of 2012 De Gruyter bought Versita and is now the world's third largest player in the Open Access branch. De Gruyter takes $2,450/1,750€ from authors for access to their Open Library.[39] Versita does not charge any money for publishing books right now, but this is limited to the first 200 books.[40] The prices of printed books will be the same as the prices of De Gruyter (p. c. Agata Morka, Product Manager, Books, 04.09.2012). Versita has interesting job offers for PhD students or postdocs: you can work with them as an Assistant Editor. The job is not paid. In exchange you get *a unique opportunity to acquire experience in*

[37] `http://www.ibi.hu-berlin.de/aktuelles/veranstaltungen/open_access/podiumsdiskussion` provides an audio recording of this event. The answer of Angelika Lex starts at 1:55:51.

[38] `http://www.springer.com/open+access/open+choice?SGWID=0-40359-0-0-0`. 22.09.2012.

[39] `http://www.degruyter.com/dg/page/16/de-gruyter-open-library`. 22.09.2012

[40] `http://versita.com/Book_Author/FAQ`. 22.09.2012.

and understanding of professional scientific publishing.[41] That is the way companies work nowadays. They get your work, you get the experience. A lot of young journalists, architects, landscapers, you name it, work that way. If you want to know where this leads to, check Section 1.2. If you want to have contact to authors, participate in professional peer reviewing, and be *Teil einer Jugendbewegung*, you can also work with us. We pay at least the same salary and are much cooler!

5.2 *Summary*

Publishing OA this way solves the problem for some of us: the readers. It does not solve the problem for the writers. Remember: we are working in underfinanced academic systems. \$3000/2000€ + 19 % taxes = a lot of money. It is two months' salary for a researcher in some of the European countries. If a German university publishes 25 articles per year that way, it could use this money to pay a programmer working on software for Open Access for a year in Germany (2000€ + 19 % × 25 = 59.500€ and 58.800€ is the amount for a full position including overheads)! In fact, the German Research Foundation (DFG) pays one million Euro per year to support this form of Open Access. I think this is nice for the researchers that are supported since their visibility is increased, but it does not solve the actual problem. As was reported in Section 2, 35 % of the money that we or the taxpayers pay will be the profit of the publishers. So, rather than financing the profit of publishing houses, the DFG and universities should finance alternative structures that do not work profit-orientated.

Finally, I want to remind the reader about economically weak countries: how should they finance such publication costs? By shifting the system from reader pays to author pays, we systematically exclude whole continents from contributing to scientific progress. This is terrible for all sciences but maybe the biggest loss for linguistics.

So, we can conclude that this form of OA publishing is not an option.

[41] `http://linguisticsnotes.wordpress.com/2012/02/20/invitation-for-assistant-editor-open-access-books-in-linguistics/`. 22.09.2012.

6 TENURE, PROMOTION, EVALUATION

A frequently asked question is whether an Open Access publication will help the author in getting tenure, whether the publication will count for promotion or in an evaluation of the research institution as they are common in Great Britain, for instance. In Great Britain the SENSE Research School uses a list of publishers that are ranked according to their quality for the evaluation of academic institutions.[42] According to this list, you get one of four possible credit points for publishing a book chapter with Peter Lang, you get zero points for publishing with Mouton De Gruyter or the Akademie Verlag. This does not correspond in any way to the importance of these publishers. While I would subtract a point from every department that publishes with Peter Lang, De Gruyter cannot be valued high enough. They run established series like *Linguistische Arbeiten* (bought from Niemeyer) and the legendary HSK series. The Akademie Verlag has the series studia grammatica, which was established in the 60's and published some of the most important books in German linguistics (Bierwisch, 1963; Kunze, 1975, 1991; Heidolph, Fläming, and Motsch, 1981; Wurzel, 1984; Stiebels, 1996). I personally own 10 books of studia grammatica, 19 books from the Linguistische Arbeiten and 10 other volumes of the various De Gruyter publishing houses, including the German version of *Le cours de linguistique générale* by Saussure and *Syntactic Structures* and *Lectures on Government and Binding* by Chomsky. I do not own any books from Peter Lang. What this is supposed to show is that the way in which research is evaluated should be changed.[43] In the meantime, however, OA initiatives have to be very careful not to get into a bad category in any of the lists that are used for evaluations.

Turning to the question of getting tenure, I guess that whether one gets tenure depends on the attitude of the tenure committee towards Open Access publishing. The list of the supporters of Open Access Linguistics Books[44] includes prominent names like Anne Abeillé, Steven Abney, Artemis Alexiadou, Johan Bos, Peter Culicover, Gisbert Fanselow, Anette Frank, Christiane Fellbaum, Charles Fillmore,

[42] `http://www.sense.nl/uploads?&func=download&fileId=855`. 14.10.2012.

[43] I do not suggest an evaluation scheme where the credits correspond to the number of books of a certain publisher on my bookshelf.

[44] `http://hpsg.fu-berlin.de/OALI/#supporters`. 27.09.2012.

Edward Gibson, Adele Goldberg, Martin Haspelmath, Hubert Haider, Ron Kaplan, Ewan Klein, Wolfgang Klein, Manfred Krifka, Gereon Müller, Steven Pinker, Friedemann Pulvermüller, Marga Reis, Ivan Sag, Stuart Shieber, Mark Steedman, Luc Steels, Tom Wasow, Dieter Wunderlich, and Annie Zaenen (to name just those with the highest h-index, but not sorted according to the h-index).[45] So, I believe that this will influence a change towards Open Access. However, those younger researchers who are insecure should go the traditional way for now, but if we manage to make the reviewing process transparent, a book publication in OALI may count more than other book publications one day. Open Access journals are already quite successful in other fields. For instance, two of the journals run by the Max Planck Society are leading their field (Living Reviews in Relativity, impact factor 17.462 best journal in "Physics, Particles & Fields" and Living Reviews in Solar Physics, impact factor 12.500, third in "Astronomy & Astrophysics").[46]

Coming back to the tenure question: whether Open Access publications make sense for an individual career also depends on the university system. As was discussed in Section 1.2, European university systems are organised very differently from the Anglo-Saxon system. While 82 % of the staff are independent researchers (professors) in the U. S., it is the other way round in Germany: 86 % work as depen-

[45] I am fully aware of the fact that the h-index is a problematic criterion when it comes to the evaluation of scientists and the German Research Foundation explicitly bans the h-index (Deutsche Forschungsgemeinschaft, 2010), but, due to some traumatic experiences at the Freie Universität, the h-index (and research evaluation in general) became one of my hobbies, and I use this index here as a measurement of citations and hence visibility.

I used Google Scholar, since the Thomson & Reuters database is unusable for linguistics. For example, Chomsky has an h-index of 16 there, whereas Friedemann Pulvermüller, working in neurolinguistics, has an h-index of 45. Chomsky's h-index in Scholarometer is 95! It's the books that make the difference! Scholarometer (formerly known as Tenurometer) should not be used in evaluations though, since they have problems with the Google API and omit citations that are marked as [citation]. This can result in up to 50 % of citations missing in search results.

For a discussion of more general problems with the hand made evaluations made by Thomson & Reuters see Rossner *et al.*, 2007.

[46] `http://www.mpg.de/5888876/impact_open_access`. 20.09.2012.

dent researchers. 74 % of the researchers are working on temporary positions. So, it is very difficult to get a permanent position and the most important thing is to stay in the system. The average age for getting tenure in Germany is 41 years. If you manage to stay in the system long enough and you keep publishing, you will get a permanent position (professorship) one day.[47] In such a system the situation is slightly different: if you publish a good book that is read a lot, the opinion of search committees will not so much depend on the question of whether the book was published Open Access or by Peter Lang, rather the impact will be the most important thing. If everybody in the search committee knows the book and if it is cited a lot, it does not matter how it was published. And the chances are high that a book gets more popular and is cited more often if it is freely available (Lawrence, 2001; Harnad and Brody, 2004; Harnad *et al.*, 2008).

7 DISCLAIMER

The question of the distribution of knowledge is independent of the question of how novels, movies, and music should be distributed. I do not share radical views about copyright. I think that artists should be paid and I have paid for all my music, either for the CD or for the downloads. Yes, I also payed for *Deichkind*'s title *Illegale Fans*. I even bought the whole album! (see Figure 4 on the next page)[48] In particular, I do not share the view of the Pirate Parties that have developed mainly in Europe but are also present in Canada and the U. S.[49] This view is also reflected in Anatol Stefanowitsch's *Open Letter to the*

[47] Thanks to Tibor Kiss, who pointed this out to me when I was frustrated.

[48] Clearly, there are things to complain about in this business as well. I remember a time in the 90's were CD prices were doubled from about 20 DM (roughly 10€) to almost 40 DM. For reasons I do not understand, the prices for AC/DC disks stayed constant. The titles that fit the description in Section 1.2 are *It's a long way to the top, if you wanna rock'n'roll* and *Ain't no fun waiting round to be a millionaire* from *Dirty Deeds Done Dirt Cheap*. Go and buy this record!

[49] The Pirate Party is represented at the government level in many German states (up to 8,9 % of the votes). One of their main political goals is the legalisation of file sharing and the prohibition of digital rights management systems. While I share their views on participation and transparency and also some of the suggestions to revise copyright regulations, I do not agree on file sharing of content without the agreement of the creators.

Rechnung an:
Stefan.Mueller@fu-berlin.de
Stefan Müller
[redacted]
DEU

Bestellnummer: MGVMZM69VF
Belegdatum: 28.02.12
Bestellung gesamt: 9,99 €
Rechnung an: Visa ....[redacted]

Artikel	Interpret	Art	Preis pro Stück
Befehl von ganz unten (Bonus Version)	Deichkind	Playlist	9,99 €
Eine Rezension schreiben Ein Problem melden			
		Bestellung gesamt:	9,99 €

Bitte bewahren Sie eine Kopie für Ihre Unterlagen auf
Die Bedingungen und Konditionen, die an diese Bestellung geknüpft sind, finden Sie weiter unten.

Figure 4: Receipt for the *Deichkind* album including *Illegale Fans*

Content Industry (Stefanowitsch, 2012) and some of the comments in his blog. Stefanowitsch suggested that artists can produce art in their spare time and that they should provide it for free and not charge others money. Here is my reaction to this view: of course, one can be an artist in one's spare time, but think of complex arrangements in music: do you believe that the Beatles or Pink Floyd could have composed their music in their spare time? Do you think that Frank Zappa should have worked as a car manufacturer or as a waiter? Or Jello Biafra as a postman?[50] Think about it: making music is not just putting a band together and playing some tunes in Joe's garage. If you play with orchestras, this takes a lot of time. Recording a song in the studio takes a lot of time. A lot of people are involved and have to be paid. How would you combine going on tour with your everyday working life? Do you think you could create photographs of the quality of Ansel Adam's pictures in your spare time? This is not possible. Believe me, I tried![51] Anatol, if you write blogs in your spare time and provide content for free, this is not your spare time! You are building a reputation (a good one, I think, if we ignore the copyright issues for a moment). This reputation adds to your value and gets you better and better paid

[50] Cf. *Stealing People's Mail* from *Fresh Fruit For Rotting Vegetables*, Dead Kennedys, Decay Music, 1980.

[51] For some pictures of linguists licenced under Creative Commons (BY-NC-ND) see LingPhot at `http://hpsg.fu-berlin.de/~stefan/Bilder/`.

jobs. Compare this with the life of a graphic designer: I used to live in an area where you can meet lots of them, they have to sublet parts of their flats to be able to survive and their living standard is comparable to the standard you have when you are employed in the humanities or below (I am talking about graphic designers with a university degree). You as a professional scientist should know that you have to use all the time you can for your research to ensure quality. If other tasks (like funny search committees that keep you busy for years) take that time, quality suffers. This is the same in art and design. Do you think a cup is designed in spare time? This is a process that takes over a year. Not everybody can do this. You can't. I can't. Do you want to have nice products? Do you want to listen to interesting music? If so, pay!

It should be clear from what has been said above that the situation in science is different: the content that is published by science publishers has already been paid (in most cases). It is the authors' free choice not to use a publisher and keep the price of their publications low, retain the copy and translation rights and offer their work for free download and indexing.

8 ACKNOWLEDGEMENTS

I thank Adam P. for inviting me to write about my views on Open Access. Writing this article made me think about all this much more intensively and I see many things clearer now.

I want to thank all those with whom I have been discussing OA issues over the past years. Special thanks go to Anatol Stefanowitsch who pointed out the existence of createspace to me. The existence of this service made me start our initiative. I thank Mary Dalrymple, Matthias Dannenberg, Christiane Fellbaum, Martin Haspelmath, Bob Levine, Frank Richter, and Stuart Shieber for discussion of and/or comments on this paper. Thanks to Bob Levine for pointers to literature concerning the h-index. I thank Monika Diecks and Remco van Capelleveen for discussing library and infrastructure issues with me. I thank Philippa Cook for the discussion of various aspects of this paper and for proof reading.

I thank Winfried Lechner and Uli Reich for providing information regarding the salaries and living costs in Greece and Brazil, respectively.

I also thank the writers of Wikipedia. While I require all my students to cite original work rather than referring to Wikipedia, I found Wikipedia extremely useful for the task of writing this article. Which conventional lexicon has an entry for the SM-4? Which 30 volume encyclopedia has entries for recent phenomena like the Pirate Parties? Thanks!

I want to also thank Brigitte Narr who runs the Stauffenburg publishing house. She is a pioneer in allowing me to have my books online as PDF files. I also thank Joachim Jacobs and Alexander Koller who supported the Stauffenburg project financially.

I thank my parents and my parents in law for discussions about the past that refreshed my memory.

Finally, I want to thank my friend Peer Höpfner for making his letters to the Kulturminister ('culture secretary') of the GDR and to the *Berliner Zeitung* and the custom form in Figure 1 available to me.

REFERENCES

Jens Allwood, Lars-Gunnar Anderson, and Östen Dahl (1973), *Logik für Linguisten*, number 8 in Romanistische Arbeitshefte, Max Niemeyer Verlag, Tübingen.

Manfred Bierwisch (1963), *Grammatik des deutschen Verbs*, number 2 in studia grammatica, Akademie Verlag, Berlin.

Hartmut Czepluch (1996), *Kasus im Deutschen und Englischen. Ein Beitrag zur Theorie des abstrakten Kasus*, number 349 in Linguistische Arbeiten, Max Niemeyer Verlag, Tübingen.

Deutsche Forschungsgemeinschaft (2010), Pressemitteilung: „Qualität statt Quantität" – DFG setzt Regeln gegen Publikationsflut in der Wissenschaft, URL `http://www.dfg.de/service/presse/pressemitteilungen/2010/pressemitteilung_nr_07/`.

Duden (1996), *Die deutsche Rechtschreibung*, volume 1, Dudenverlag, Mannheim, Leipzig, Wien, Zürich, 21 edition.

Michael Gamon and Tom Reutter (1997), The Analysis of German Separable Prefix Verbs in the Microsoft Natural Language Processing System, Technical Report MSR-TR-97-15, Microsoft Research, Redmond, WA, URL `http://research.microsoft.com/apps/pubs/default.aspx?id=69549`.

Stevan Harnad and Tim Brody (2004), Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, *D-Lib Magazine*, URL `http://www.dlib.org/dlib/june04/harnad/06harnad.html`.

Stevan HARNAD, Tim BRODY, François VALLIÈRES, Les CARR, Steve HITCHCOCK, Yves GINGRAS, Charles OPPENHEIM, Chawki HAJJEM, and Eberhard R. HILF (2008), The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update, *Serials review*, 34(1):36–40.

Karl Erich HEIDOLPH, Walter FLÄMING, and Walter MOTSCH, editors (1981), *Grundzüge einer deutschen Grammatik*, Akademie Verlag, Berlin – Hauptstadt der DDR.

Peer HÖPFNER (1985), Reader's Letter to *Berliner Zeitung*, URL `http://hpsg.fu-berlin.de/~stefan/Pub/OA-JLM/1985-Leserbrief-Berliner_Zeitung.pdf`.

Joachim JACOBS, Arnim VON STECHOW, Wolfgang STERNEFELD, and Theo VENNEMANN, editors (1993), *Syntax – Ein internationales Handbuch zeitgenössischer Forschung*, volume 9.1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, Walter de Gruyter Verlag, Berlin/New York, NY.

Joachim JACOBS, Arnim VON STECHOW, Wolfgang STERNEFELD, and Theo VENNEMANN, editors (1995), *Syntax – Ein internationales Handbuch zeitgenössischer Forschung*, volume 9.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, Walter de Gruyter Verlag, Berlin/New York, NY.

Brian W. KERNIGHAN and Dennis M. RITCHIE (1978), *The C Programming Language*, Prentice Hall, Englewood Cliffs, NJ.

Reinhard KRECKEL, editor (2008), *Zwischen Promotion und Professur. Das wissenschaftliche Personal in Deutschland im Vergleich mit Frankreich, Großbritannien, USA, Schweden, den Niederlanden, Österreich und der Schweiz*, Akademische Verlagsanstalt, Leipzig.

Jürgen KUCZYNSKI (1983), *Dialog mit meinem Urenkel – Neunzehn Briefe und ein Tagebuch*, Aufbau Verlag, Berlin/Weimar.

Jürgen KUNZE (1975), *Abhängigkeitsgrammatik*, number 12 in studia grammatica, Akademie Verlag, Berlin.

Jürgen KUNZE (1991), *Kasusrelationen und semantische Emphase*, studia grammatica XXXII, Akademie Verlag, Berlin.

Steve LAWRENCE (2001), Free Online Availability Substantially Increases a Paper's Impact, *Nature*, 441:521.

Stefan MÜLLER (2008), *Head-Driven Phrase Structure Grammar: Eine Einführung*, number 17 in Stauffenburg Einführungen, Stauffenburg Verlag, Tübingen, 2 edition, URL `http://hpsg.fu-berlin.de/~stefan/Pub/hpsg-lehrbuch.html`.

Stefan MÜLLER (2010), *Grammatiktheorie*, number 20 in Stauffenburg Einführungen, Stauffenburg Verlag, Tübingen, URL `http://hpsg.fu-berlin.de/~stefan/Pub/grammatiktheorie.html`.

Stefan MÜLLER (Submitted), Unifying Everything, URL `http://hpsg.fu-berlin.de/~stefan/Pub/unifying.html`, ms, Freie Universität Berlin.

Stefan MÜLLER, Pollet SAMVELIAN, and Olivier BONAMI (In Preparation), *Persian in Head-Driven Phrase Structure Grammar*, URL `http://hpsg.fu-berlin.de/~stefan/Pub/persian.html`.

Renate PASCH, Ursula BRAUSSE, Eva BREINDL, and Ulrich Herrmann WASSNER (2003), *Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln)*, number 9 in Schriften des Instituts für deutsche Sprache, Walter de Gruyter, Berlin, New York.

Geoffrey K. PULLUM (1984), Stalking the Perfect Journal, *Natural Language and Linguistic Theory*, 2(2):261–267.

Jon RADOFF (2011), *Game On: Energize Your Business with Social Media Games*, Wiley.

Mike ROSSNER, Heather VAN EPPS, and Emma HILL (2007), Show Me the Data, *The Journal of Cell Biology*, 179(6):1091–1092.

Anatol STEFANOWITSCH (2012), Offener Brief an die Contentindustrie, URL `http://www.scilogs.de/wblogs/blog/sprachlog/sprachwandel/2012-04-06/offener-brief-an-die-contentindustrie/page/4`.

Barbara STIEBELS (1996), *Lexikalische Argumente und Adjunkte: Zum semantischen Beitrag verbaler Präfixe und Partikeln*, number 39 in studia grammatica, Akademie Verlag, Berlin.

Wolfgang WURZEL (1984), *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung*, number 21 in studia grammatica, Akademie Verlag, Berlin.

Slovak Morphosyntactic Tagset

Radovan Garabík and Mária Šimková
Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences,
Bratislava, Slovakia

ABSTRACT

Morphological annotation constitutes essential, very useful and very common linguistic information presented in corpora, especially for highly inflectional languages. The morphological tagset used in the Slovak National Corpus has been designed with several goals in mind – the tags are compact and easily human-readable, without sacrificing their informational contents. The tags consist of ASCII letters, numbers and several other characters. In general, they have a variable number of symbols, but their order is obligatory, and each category or specific feature is assigned a particular character, which can be shared among several parts of speech. The tagset is highly functional and pragmatic, although some allowances had to be made to accommodate the traditional analysis of Slovak morphology and part of speech categories.

Keywords:
Slovak language, corpus, tagset, morphology, part of speech, grammatical categories

1 INTRODUCTION

Morphological annotation constitutes fundamental and very common linguistic information found in corpora, especially for inflectional languages. It comprises the part of speech categorisation of lemmas and morphological characterisation of a word (token).

It is usually preceded by the process of lemmatisation (an assignment of the basic form to a particular lexeme). Since Slovak belongs to a family of highly inflectional languages, a morphological annotation is not a simple and straightforward process. Currently, the process of morphological analysis of such languages is often performed in two steps; the first one is the analysis itself (assigning to each of the words a list of possible combinations of lemmas and morphological tags), and

the second one is disambiguation, picking up one (correct, if possible) lemma-tag combination. The analysis itself is often nothing other than selecting the entries from a database of inflected wordforms (with an additional step of guessing lemmas and/or tags for out-of-dictionary words). The second step is often performed using statistical methods, requiring training on manually annotated corpora.

In the Slovak National Corpus (SNK), morphological annotation and lemmatisation occurs prominently in two places:

- manual morphological annotation and lemmatisation of the *r-mak* subcorpus
- automatic morphological annotation of the whole corpus (and other relevant corpora and subcorpora)

The *r-mak* subcorpus is a manually lemmatised and annotaded corpus of 1.2 million tokens (punctuation included). The progress from version 3.0 (released in 2008) to 4.0 (released in 2012) did not encompass any new texts; rather the existing annotations have been semi-automatically proofread and corrected, several duplicities have been identified and removed, the revision of the tagset has been applied where necessary, and a new, more consistent sentence segmentation has been introduced.

Thus, the end users of the corpus (corpora) meet the analysis while using the corpus, either when entering more complex queries or when displaying grammatical categories of the results. In this article, we describe the tagset in detail, including the motivation behind some design choices.

As the Slovak National Corpus at its inception in 2002 was primarily aimed at linguistic (mainly lexicographical) use, the morphological annotation and tagset were created with this in mind – the design of the tagset was based on the formalised Slovak language morphology (Páleš, 1994; Benko *et al.*, 1998), traditional grammar description (Dvonč *et al.*, 1966) and other similar tagsets of related inflectional languages (Hajič and Vidová-Hladká, 1997; Džeroski *et al.*, 2000; Hajič, 2000; Dębowski, 2001). Tokenisation, lemmatisation and the principles of morphological annotation used in manual tagging of the *r-mak* corpus are described in the user guide (Garabík *et al.*, 2004). The tagset is used in the morphological database of the Slovak language,

covering (at the time of writing) more than 97 thousand lemmas and about 3.2 million inflected and tagged entries (Garabík, 2006).

The new revision of the tagset and some of the principles occurred in 2012. It did not introduce any new tags, but rather clarified many borderlike cases and the classification of many words has been re-evaluated (based on actual corpus evidence and inconsistencies introduced therein). This article presents some of the reasoning behind the decisions.

All the examples used in this article are based on actual text occurrences in the Slovak National Corpus.

2 TOKENISATION

The tagset is designed to cover morphology of the smallest possible units – this governs the tokenisation principles. Most notably, there are no multi-word tokens; each constituent of such an element is a separate token. This includes also hyphenated words – expressions like *slovensko-poľský* (Slovak-Polish) will be tokenised as three tokens: *slovensko*, - (i. e. the hyphen), *poľský*. The advantage of this is a clear and unambiguous approach to the tokenisation, but as a main disadvantage, we lose a reasonable way of dealing with multiword expressions, and even have to introduce a special morphology tag to mark constituents of such expressions.

3 OTHER SLOVAK LANGUAGE TAGSETS IN USE

3.1 *Tagset developed at the Institute of Formal and Applied Linguistics*

This tagset is an adaptation of the Czech language tagset developed at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague (Hajič, 2004). It is a positional tagset of fixed length, each tag containing 15 ASCII characters. Each position encodes one (grammar) category, and some of the positions are empty (13, 14). The first position encodes part of speech.

A distinguishing feature of this tagset is a very detailed description of the part of speech subdivision (position 2): e.g. there are 16 different types of numerals, 21 types of pronouns.

Most notably, the tagset does not encode verbal aspect (an omission inherited from the Czech tagset). With some effort and a database of perfective and imperfective verbs, it can be inferred from the lemma – indeed, for the Czech language this has been done in the extended version used in the Czech National Corpus[1].

3.2 *Majka/Ajka*

Majka is a Czech language morphological analyser developed at the Faculty of Informatics, Masaryk University in Brno (Šmerk, 2010), a reimplementation of the previous analyser *ajka* (Sedláček, 2001). The tagset has been carried over to the Slovak version of *ajka*. It is an attributive tagset, with one-letter codes for the attributes and one-letter codes for the values. The codes for the values can be reused across attributes – the tags are of unequal length (although a rather important feature is that the value assignment does not depend on a part of speech).

3.3 *Multext East*

The EC INCO-Copernicus project MULTEXT-East Multilingual Text Tools and Corpora for Central and Eastern European Languages (Dimitrova *et al.*, 1998) developed language resources for six Central and Eastern European (CEE) languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovene, as well as English.

Slovak language morphology specification compatible with the MULTEXT-East (MTE) tagset was not part of the original Multext East specification – it has been developed separately at the Ľ. Štúr Institute of Linguistics (Garabík, 2011). The tagset follows MTE principles and tries to be compatible with the other MTE language tagsets, and especially with Czech (some of the design features were directly inspired by solutions deployed in the Czech MTE tagset). The tagset has been influenced by the Slovak National Corpus tagset described herein – one of the design goals was to make an automatic conversion from the SNK tagset into the MTE not too difficult. This even meant removing some features from the MTE tagset if they could not be inferred from the information about SNK tag and lemma (e.g. the verb *byť* (to be) is always referred to as Type=c(copula)).

[1] `http://korpus.cz`

This tagset is used exclusively in the scope of the MTE project and related research (Garabík *et al.*, 2009).

4 GENERAL PRINCIPLES

While the attributive versus positional tagset dispute is not a very important one (after all, a tagset is just a representation of grammar features and a mapping from one representation into a different one provides no inherent insights), we have to realise that a tagset is something that will be with the users for some time, especially if we are designing a tagset to be used in a big ('national') corpus. Both attributive and positional tagsets have their advantages and disadvantages – the positional system is often opaque to the user; if the number of positions is big enough, it is difficult to find out which position is which without counting the positions. Attributive tagsets tend to be even longer, because each value has to be accompanied by its attribute; but if the attribute abbreviations are selected sensibly, the users can decode the meaning at a glance.

In designing our tagset, we tried to extract the best of the two words while keeping the disadvantages to a minimum. One of the most important design decisions is to keep the meaning of codes unambiguous – one letter should correspond to one value *only*, even across parts of speech. The only exception is the paradigm category, which reuses the part-of-speech code. We try to assign mnemonic, easily-remembered codes familiar from common Slovak education and linguistic environment whenever possible. The tags are of unequal length, but most tags follow the same structure for the same inflectional paradigm (not the part of speech category). These principles make it easy to test grammar categories in software. Checking the part of speech category could be expressed in a Python-like pseudolanguage as[2]:

```
if tag[0]=='S': # noun
    # proceed with the noun
```

and checking for the value of the grammar category can be as easy as:

[2] We are counting the positions from zero.

```
if '4' in tag: # accusative
    # proceed with the accusative
```

4.1 *Tag Structure*

As the part of speech information is often the most important, it is encoded in the first[3] position. The second position usually marks an inflectional paradigm for words that do have this category. The code for the position repeats that of corresponding parts of speech, e.g. `SS`… stands for a noun with a noun-like inflectional paradigm, and `PS`… for a pronoun with a noun-like inflectional paradigm. First we describe the symbols and corresponding grammar categories, then we discuss motivation and principles behind several choices for individual part of speech categories.

The tag can be optionally followed by a marker separated by a colon. The marker is used to denote proper names (symbol `:r`) and erroneous words (symbol `:q`). The definition of 'erroneous' is strictly limited to typos and errors caused by text conversion – substandard words, dialectical words and frowned-upon expressions are not tagged as erroneous (unless they contain a typo). The symbols can be combined for an erroneous proper name (`:rq`). In the following example, the surname *Kirscher*/`SUfs7:rq` should have been *Kirschner*/`SUfs7:r`.

(1) *duetu* *s* *Janou* *Kirscher*
`SSis2` `Eu7` `SSfs7:r` `SUfs7:rq`
'a duet with Jana Kirsch[n]er'

5 MAIN GRAMMATICAL CATEGORIES

5.1 *Paradigm*

Slovak language exhibits certain discrepancy between morphological and syntactic behaviour of words (a behaviour shared with other inflected languages). This is reflected in various ways in traditional grammar descriptions, usually by classifying the word to be of a part of speech category corresponding to its morphological class, but acknowledging that the word "behaves as if it were of a different category". Such an ambiguity in description does not have a place in de-

[3] In the following, we count and describe the positions starting with 1, as is customary in many human languages.

signing a morphological tagset, unless we introduce a special category for such ambiguous words, which is something that we were trying to avoid. We introduced the 'paradigm' category, which describes the morphological (inflectional) behaviour of the word. It is in fact a conflation of two different ideas – the inflectional pattern of a different part of speech present in another part of speech, but also describes several other, non-mainstream inflectional patterns.

The paradigm category is specified for nouns, adjectives, pronouns and numerals. The symbol is equal to that of the part of speech category the paradigm follows, with some additional types.

We recognise the following paradigms:

Substantive (symbol `S`) – used for nouns, pronouns and numerals.

Adjectival (symbol `A`) – used for nouns, adjectives, pronouns and numerals.

Pronominal (symbol `P`) – used for pronouns.

Numeral (symbol `N`) – used for numerals.

Adverbial (symbol `D`) – used for pronouns and numerals.

Mixed (symbol `F`) – used for nouns, adjectives, pronouns and numerals. This paradigm is used for words that do not clearly follow one inflectional pattern but instead exhibit features from two or more morphological parts of speech.

Incomplete (symbol `U`) – used for nouns, adjectives, pronouns and numerals. This is used in a case where the word does not exploit all the morphological inflections, typically an uninflected noun or an adjective, 3rd person possessive pronouns and (some) cardinal numerals.[4]

5.2 *Grammatical Number*

There are two grammatical numbers in Slovak, singular and plural. Of the old Slavic dual there are only some traces left, but they are not in contrast with singular and plural (unlike e.g. in Czech, where the dual still manifests itself in several nouns – body parts – in the instrumental). We use `s` as the symbol for the singular and `p` for the plural; there are no provisions for marking pluralia and singularia tantum.

[4] Not to be confused with a 'partial' (or 'incomplete') paradigm where a part of the paradigm is missing, such as pluralia tantum.

5.3 *Grammatical Gender*

In Slovak, 3 traditional genders are recognised, but in our analysis we split the masculine animate and masculine inanimate to get 4 different genders: masculine animate – `m`, masculine inanimate – `i`, feminine – `f`, neuter – `n`. There are two more 'genders' marked in the tagset, general – `h` and undefined – `o`. These are used as a conflation of other genders in cases where disambiguating them would be impractical or directly impossible. Personal pronouns use the `h` (general) symbol for everything except the third person ones (*on, ona, ono, oni, ony*). In the 1st or 2nd person, the pronouns could be reasonably assigned a gender only in the presence of an adjective or a verb in conditional or past tense – in a typical sentence with a verb in the indicative form it is impossible. Verbs use the `h` for the L-participle plural in the first and second person (in agreement with corresponding personal pronouns, which is also marked with the 'general' gender) and the `o` (undefined) for the third person if the verb covers several genders at once – e.g. the following example has the verb *kričali* (yelled) tagged with the undefined gender, because there are two subjects in the sentence – *muž* is masculine, but *žena* is feminine.

(2) *muž* *a* *žena* *na* *seba* *kričali*
`SSms1 O SSfs1 Eu4 PPhs4 VLepco+`
'[the] man and woman yelled at each other'

5.4 *Case*

Slovak distinguishes 6 cases, the locative case being obligatorily prepositional and the nominative obligatorily non-prepositional. We fully realise there is no separate vocative case described by traditional grammars in the contemporary system of Slovak language morphology. What we called a "vocative" in this article is in fact a syntactical role of a noun when used for addressing someone, a role that is only sometimes realised morphologically and in most of the cases is identical with the form of the nominative case. The exceptions exist in the case of several nouns (fossilised forms of old Slavic vocative) such as *bože, pane, priateľu, človeče* … (God, Sir, friend, man) and (substandard usage of) some proper names and interpersonal relationship terms – *Zuzi, babi, oci, mami, tati, šéfe* … (Susan, grandma, dad, mum,

dad, boss). If this article were about Russian, we would use the term "new vocative" here (see e.g. Comtet, 1997).

The cases were traditionally numbered (starting with elementary and secondary school syllabi) and Slovak linguistic and general audience is familiar with case numbers. The numbering went 1-nominative, 2-genitive, 3-dative, 4-accusative, 6-locative, 7-instrumental. We decided to retain this numbering in our tagset, so the numbers 1 through 7 reflect these cases (with the number 5 for the vocative).

5.5 *Degree of Comparison*

Slovak has three degrees of comparison: positive, comparative and superlative. The degree is defined only for adjectives, participles and adverbs, and we assigned to it the symbols `x` for positive, `y` for comparative and `z` for the superlative, for all these three parts of speech.

6 PART OF SPEECH CATEGORIES

6.1 *Noun*

The noun tag is of a fixed length of 5 positions:

Position	Possible values	Description
1	S	part of speech tag
2	SAFU	paradigm
3	mifn	gender
4	sp	number
5	1234567	case

The `S` paradigm stands for 'normal' nouns with a full, substantive-like morphology. The `A` (adjectival) paradigm stands for substantivised adjectives or participles. These are often distinguished by proper adjectives only by their semantic role and there often exists an identical adjective or a participle as well. Examples include *obžalovaný* (accused, a passive participle of *obžalovať*), *cestujúci* (traveller, an active participle of *cestovať*), *zelený* (a member of the Green movement; adjective when it is a colour term). The `U` paradigm is used for uninflected nouns – the same form in all the cases and numbers, either completely domesticated loanwords like *kupé*/`SUns1`, *finále*/`SUns1`, or loanwords like *whisky*/`SUfs1`, *miss*/`SUfs1`, or several native substantivised short phrases

or words like *skaderuka*/sUms1-*skadenoha*/sUms1. It is also used for letters (of the alphabet) when used as nouns (e.g. as in the sentence *Od*/Eu2 *A*/sUns2 *po*/Eu4 *Z*/sUns4 *./z*). The F paradigm is used for nouns which combine different inflectional paradigms – e.g. *princezná*/sFfs1 inflects like an adjective, with the exception of genitive (*princezien*/sFfp2), locative (*princeznách*/sFfp6) and instrumental (*princeznami*/sFfp7) plural.

Many animal names in Slovak are masculine animate in the singular, but depending on the familiarity and the degree of anthropomorphisation, the plural can be either animate or inanimate. The rule of thumb is: the higher the organism, the more animateness it shows. People are always animate; *pes* (dog) is animate in the singular and can be both in the plural; *jeleň* (deer) is mostly inanimate in the plural (but can be sometimes animate); *bacil* (germ) is inanimate in the singular and plural, but there are cases of animate singular appearing[5], and *stroj* (machine) is inanimate without exception. The animateness is sometimes used as a semantic disambiguator – *android* is mostly animate when it is a humanoid robot, but mostly inanimate when it is an operating system.

This is reflected in the morphological database – there are lexemes that are masculine animate in the singular and masculine inanimate in the plural, or the plural entries have two variants (inanimate and animate). There are even some singular cases, e.g. *knieža* (duke) is neuter, with the exception of nominative singular, which can be also masculine (the form is the same, but the gender governs adjective and verb agreement), so the tags for *knieža* could be both SSms1 and SSns1.

6.2 *Adjective*

The adjective tag is of a fixed length of 6 positions:

Position	Possible values	Description
1	A	part of speech tag
2	AFU	paradigm
3	mifn	gender congruence
4	sp	number congruence
5	1234567	case congruence
6	xyz	degree of comparison

[5] Not all of the examples cited are ‘correct’ by official language rules and dictionaries, they however have non-negligible corpus evidence.

The `U` paradigm stays for indeclinable adjectives. These include a rather exceptional case: three fossilised short forms *hoden*/`AUms1x` (worth), *vinen*/`AUms1x` (guilty) and *dlžen*/`AUms1x` (indebted). These forms also have a different syntactical usage from their regular counterparts. occur only in nominative singular and that is the only entry in the database – the regular long forms *dlžný*/`AAms1x`, *vinný*/`AAms1x`, *hodný*/`AAms1x` are separate lexemes (and the short forms have already shifted semantically). Other indeclinable adjectives are e.g. *nanič*, *akurát*[6], *rád*, special form *naj* (superlative prefix, when used standalone in an adjectival function), and many loanwords at various level of domestication – *super, fajn, hurá, bianko, nealko* …

The `F` paradigm marks possessives (which are considered to be adjectives in traditional Slovak grammars) – e.g. *tetin*/`AFis1` *hlas*/`SSis1`, *Sapkowského*/`AFfp1x` *knihy*/`SSfp1`. The gender in the tag agrees with the gender of the possessed; the gender of the possessor is not marked. It is also used for the special adjective *nesvoj*/`AFms1x`, which is morphologically identical with possessives (derived from a possessive pronoun *svoj*/`PFms1`).

In ambiguous cases (where even traditional grammar descriptions admit that a decision cannot be taken unambiguously) (e.g. *nepočujúci*), we sorted the words according to their attested usage in the corpus – the word was classified as an adjective only if there was a significant percentage of its occurences in adjectival positions (i. e. modifying a noun), disregarding intentionally defective or metalanguage usage. Such decisions have been consulted with the Short Dictionary of the Slovak Language (Považaj, 2003), but preferring the actual corpus evidence.

6.3 *Pronoun*

The structure of the pronoun tags depends on the pronoun inflectional paradigm (roughly, the tag structure follows that of the corresponding part of speech of the paradigm type).

[6] Note that *nanič* and *akurát* can be adverbs as well, in adverbial constructions.

Position	Possible values			Description
1	P	P	P	part of speech tag
2	SAFU	P	F	paradigm
3	mifn	h	min	gender
4	sp	sp	s	number
5	1234567	1234567	24	case
6			g	agglutinated

The pronouns are split into three subclasses, according to the paradigm position. The table above captures the three possible combinations of values in the 'Possible values' columns and can be viewed as a concise combination of three different tables, one for the `S, A, F, U` paradigms, the second one for the `P` paradigm and the third one for the `F` paradigm (which is longer by one position, the 'agglutinated' value.

The paradigm `A` is used for adjective-like pronouns: *aký, ktorá, inakšie, samý*.

`F` is used for pronouns that do not have clearly separated morphosyntactical paradigm, typically possessives, e.g. *môj, tvoj, svoj, tento, táto, toto*, and basic personal pronouns *ja, ty, my, vy, seba*.

`U` is used for pronouns that do not decline. These are 3rd person possessives *jeho* (his, its), *jej* (hers) and *ich* (theirs), and *koľko, toľko, bárskoľko, hockoľko* ….

The symbol `g` marks agglutinating of preposition and pronoun – in majority of pronoun tags it does not occur and the tag is then 5 characters long. It appears in pronouns like *preňho*/`PFms4g`, *doň*/`PFms2g` (which are fusions of *pre*/`Eu4` *neho*/`PFms4`, *do*/`Eu2` *neho*/`PFns2`. These pronouns are lemmatised as *pre_on*, *do_ono* (i.e. the combination of a preposition and a pronoun, joined by an underscore). The only existing tags with the agglutinating symbols are `PFms2g`, `PFis2g`. `PFns2g`, `PFms4g`, `PFis4g`. `PFns4g` (i.e. only genitive or accusative singular, non-feminine gender). These agglutinations are traditionally described as pronouns, which was the main reason for including them in this category.[7]

The uninflected adverbial pronouns are tagged with the tag `PD`: *ako, tak, prečo, načo*.

[7] The alternative would be to tokenise them as two different tokens; this would however complicate the tokenisation phase.

6.4 *Numeral*

Position	Possible values	Description
1	N	part of speech tag
2	SANFU	paradigm
3	mifn	gender
4	sp	number
5	1234567	case

The paradigm follows the morphology of the numeral, but since the morphology reflects the numeral type, it is also useful for determination of the type.

The `N` paradigm describes small cardinal numerals (2, 3, 4). These are inflected and always in the plural. For the numeral 2 all the genders have separate inflections, and for the cardinality of 3 and 4 the masculine animate gender is in contrast to other genders in the nominative and accusative.

The tag `S` is used for other numerals that inflect like nouns – fractions like *tretina*, *štvrtina*, huge cardinals like *milión*, *septilión* and the word *raz* (once).

The `F` paradigm is used for the cardinal number 1. The numeral is inflected for gender, case and number (the plural is used for group numerals).

The `U` paradigm is used for other cardinal numbers (5, 6, 7, …).

The `A` paradigm describes numerals with adjective-like inflection – primarily ordinal numerals, but also several indefinite ones like *mnohý*`/NAms1`.

The tag `ND` (not inflected, without any other grammar categories) is used for adverbial numerals, e.g. *neraz*, *prvýkrát*.

Not only are the inflectional patterns of several classes of these numerals (noun-like, adjective-like) identical to the corresponding parts of speech, but their syntactic behaviour is also equivalent. In this regard, the usage of the `N` tag differs from other parts of speech, because it encodes also their semantic role. This behaviour was retained from the traditional grammars and the description present in the Short Dictionary of Slovak Language[8].

[8] Apart from the word *polovica* ([one] half), which is considered to be described erroneously as a noun in the dictionary

6.5 *Verb*

The tag for verbs is probably the most complicated out of the whole tagset. It does not have a fixed length; the length is determined by the second position, which in this case does not mark the paradigm, but the form of the verb.

We do not adhere strictly to established grammar categories, but follow the verbal form instead. This is the reason we do not mark the tense as such.

Each tag is however at least 4 positions long, and these four positions have a fixed meaning. The third position marks aspect – there are three possible values: `d` for the perfective aspect, `j` for the imperfective one and `e` for the ambivalent verbs. The ambivalent aspect actually means the perfective and imperfective verb forms are identical (but e.g. they form the future tense differently, according to their aspect). Since they are identical in their morphology and we follow strictly formal morphological criteria, we do not try to disambiguate them.

The last position marks positiveness/negativeness – we use the plus sign + for positive verbs and the minus sign - for negated ones. The negation of Slovak verbs is formed with the *ne-* prefix (e.g. *kompiluj* will be negated as *nekompiluj*) invariably in all the conjugated forms. There are some verbs that lack the negated form (e.g. *nenávidieť*/`VIe+`, although some corpus evidence exists, it points out to metalanguage usage or puns) and for these the negated tag does not appear. The only exception is the indicative of the verb *byť* (to be), which is negated by a separate particle *nie*, written separately (this is just an orthography quirk). These cases are tokenised as two separate tokens, with the first one tagged as a particle and the second one as a (positive) verb:

(3) *jazyk* *nie* *je* *usporiadaný*
`SSis1` `T` `VKesc+` `Gtis1x`
'language is not ordered'

This does not explicitly contain information about the 'negativeness', but marking it in any other way would introduce other inconsistencies (e.g. marking the negativeness of a morphologically positive verb or marking the particle as a verb).

We describe the tags sorted by their length, going from the longest (the most complicated) to the shortest one.

Position	Possible values	Description	Position	Possible values	Description
1	V	part of speech tag	1	V	part of speech tag
2	L	form (L-participle)	2	KMB	form
3	dej	aspect	3	dej	aspect
4	sp	number	4	sp	number
5	abc	person	5	abc	person
6	mifnho	gender	6	+ −	negation
7	+ −	negation			

The L-participle is used to form past tense(s) and conditionals. Sharing some features with participles, it distinguishes number and gender, and these appear in the tag, making it 7 positions long. Two extra genders – `h` and `o` – were described in Section 5.3.

The indicative (tag `K`) is used to form a present tense for imperfective verbs, and a future tense for perfective ones. For ambivalent-aspect verbs, the indicative form can mean either a present or a future tense, depending on the meaning of the verb. The indicative of the verb *byť* is also used to form the past tense (together with the L-participle). We do not distinguish this auxiliary usage of *byť* from the copula.

The imperative (tag `M`) is also marked for number and person. In the singular, only the second person is possible; in the plural, imperatives can have both the second and the first (inclusive) person.

The future (`B`) is mostly used for the future form of the auxiliary verb *byť*, which is used to form the future tense of imperfective verbs (together with their infinitive), and for the simple future of the copula *byť*. It is also used for a small class of verbs of movement, which form the future tense with the prefix *po-*.

Position	Possible values	Description
1	V	part of speech tag
2	IH	form
3	dej	aspect
4	+ −	negation

The transgressive (symbol H) in Slovak is morphologically derived from the 3rd person plural indicative, usually by adding the suffix *-c*. It has only one form and in contrast to Czech, it is not distinguished either for number or gender, and there is only a present transgressive. The transgressive is marked just with the aspect and negation – *čítajúc*/VHj+, *neuznajúc*/VHd-.

The infinitive (symbol I) have just one form and is also marked only with the aspect and negation.

6.6 *Participle*

There are two different classes of participles in Slovak – active and passive (the L-participle has been discussed in the section on Verbs). The participles exhibit a strong adjective-like morphological behaviour, up to being inflected for a degree of comparison. Their classification as verbs or adjectives is a perennial problem in many languages, and either way leads to some unsatisfactory behaviour. In the Russian Multext East tagset (Sharoff *et al.*, 2008), the participles belong to the Verb category and as such have the case attribute – this has been facilitated by the Multext East formal appearance – the 'case' position is always present, it is just left undefined for the verbs, and it is very easy to reuse it for participles. It is not clear if this coincidence was decisive in categorising the participles or not.

On the other hand, in the Czech Multext East tagset (Dimitrova *et al.*, 1998) the participles are not distinguished in any way from adjectives – they have the 'qualificative adjective' attribute.

We consider the participles to be a separate part of speech class, not a declined form of verbs – while definitely possible, this would lead up to some singular categorisation, e.g. verbs with case. The participles are functionally very similar to adjectives, and indeed many an adjective has originated as a participle of which the source verb is no longer in the language. Sometimes the boundary between participles and adjectives is rather unclear – in ambiguous cases, we conformed to the Short Dictionary of the Slovak Language, which is an arbitrary solution, but probably the best one, given the status of the dictionary. For cases not clearly stated in the dictionary we leaned towards the participles if there existed (at least formally) the source verb.

The passive participle is not distinguished for tense, but formally the active participles are separated into present active and past active ones. The present active participle is commonly found in standard Slovak, but the past active participle is dead for all practical reasons in both literary and standard Slovak – there are only 7 occurrences of the form in the manually annotated corpus, 6 of them from the same document (a treatise on liturgic history).

For this reason, we decided not to introduce any special category separating past and present active participles and use the same tags for both of them. However, we differentiate between passive and active participles.

Position	Possible values	Description
1	G	part of speech tag
2	kt	type
3	mifn	gender congruence
4	sp	number congruence
5	1234567	case congruence
6	xyz	degree of comparison

The type of the participle is marked by the second symbol in the tag – `k` for active, `t` for passive ones. The rest of the symbols follows the symbols for the adjective. Participles can also have a degree of comparison, even if the comparatives and superlatives occur rather rarely.

6.7 *Adverb*

The tag for an adverb is invariably two letters long:

Position	Possible values	Description
1	D	part of speech tag
2	xyz	degree of comparison

The degree of comparison is always specified, even if neither comparative nor superlative exists for the adverb (e.g. *neveľmi*/Dx). While we could consider marking irrelevant degrees of comparison (as opposed to positive ones), for the sake of consistency we decided to unify these two cases – this also saves us from having to invent excuses for claiming that a given word has irrelevant degree (according to traditional grammars), even if corpus evidence suggests otherwise.

6.8 *Preposition*

Position	Possible values	Description
1	E	part of speech tag
2	uv	vocalisation
3	23467	case (valency)

Some of the prepositions ending with a consonant exhibit vocalisation – a vowel is appended after the preposition ending with a consonant in certain cases, mostly in non-syllabic preposition, if the next word begins with a consonant of the similar class as the last consonant of the preposition, e.g. if both consonants are sibilants, or both consonants are velar stops, or both consonants are alveolar stops – *k/*$_{Eu3}$ *domu/*$_{SSis3}$ (to [the] house); *ku/*$_{Ev3}$ *korpusu/*$_{SSis3}$ (to [the] corpus), or in some other fixed epressions – *bez/*$_{Eu2}$ *strachu/*$_{SSis2}$ (fearless); *bezo/*$_{Ev2}$ *mňa/*$_{PPhs2}$ (without me).

We mark the vocalised prepositions with the symbol v at the second position, the non-vocalised ones are marked with the symbol u. The lemma of the vocalised prepositions is the non-vocalised form. The third position of the tag encodes the case the preposition binds with (nominative and vocative are not present, according to existing grammar theories).

Compound prepositions are analysed as a sequence of constituents, if possible – e.g. *s/*$_{Eu7}$ *ohľadom/*$_{SSis7}$ *na/*$_{Eu4}$ is tagged as a preposition, a noun and a preposition. There is a sizeable amount of fossilised noun or verb forms that have become prepositions, and these are marked as prepositions (*postupom*, *doprostriedku*, *končiac*). There is often a homonymy with adverbised fossilised forms as well. This makes the class of prepositions less closed and unambiguous than we would like.

In Slovak, no preposition that binds the nominative exists – the reason is that nominative is obligatory non-prepositional and the reason for the nominative to be obligatory non-prepositional is that no prepositions binding with the nominative exist. This circular reasoning is generally accepted in traditional linguistic circles, and the loans *à*, *à la* (often domesticated as *á*, *á la* or *a la*) are not considered to be prepositions, but particles instead.

In our tagset, we did not dare to break this tradition, and *à la* will be tagged as two tokens, a residual *à/*$_{Q}$ and the particle *la/*$_{T}$ (see below for the description of the tags).

(4) *šat* *à* *la* *Zajac*
SSis1 Q T SSms1:r
'clothing à la Zajac'

6.9 *Other Categories*

Conditional morpheme *by* has a special tag Y. This standalone morpheme is used to form conditionals (with the L-participle), but it can also form multiword prepositions or conjunctions (*nie že by*). However, such multiword prepositions are followed by the L-participle and semantically introduce conditional clauses, therefore it is easier to consider the *by* to be part of the conditional in these circumstances as well. We decided to tag the *by* in all these cases with the same tag. The homonymous poetic conjunction *by* (an abbreviation of *aby*) is also tagged by the Y, which is an inconsistency with other conjunctions, but it is justified by the highly poetic (and therefore rather infrequent) nature of the word and the need to keep the ambiguity low.

Since the morpheme fused with some other functional words, the symbol Y is also used as a second symbol in several other part-of-speech tags, to denote the fusion.

Punctuation characters have their own one-letter long tag Z. Lemmas of the punctuation characters are 'normalised' – various types of quotes are lemmatised as straight quotes U+0022 QUOTATION MARK, hyphens and dashes as U+002D HYPHEN-MINUS.

Conjunctions can be either simple, having the one-letter tag O (*a, aj, alebo, než...*), or they can contain a fused conditional morpheme *by*, with a two-letter tag OY (*aby, keby, akoby, niežeby, žeby, sťaby, čoby, nietoby, nietožeby*).

Particles are tagged with the tag *T*. Similar to conjunctions, some of the particles contain fused conditional morpheme *by* (*čoby, kiežby, žeby*) and are tagged with TY.

Abbreviations are tagged with W. We do not distinguish between abbreviations and acronyms, and we do not assign any other grammar categories to it (even if the abbreviated words have them), e.g. in

(5) *odborní pracovníci SNK podali obraz*
AAmp1x SSmp1 W VLdpcm+ SSis4
'SNK professionals gave impression'

the *SNK* is an abbreviation, even if it can be thought of as a noun in genitive singular. As an artifact of our tokenisation, if there is a trailing dot, it is not a part of the token, but a separate token with the punctuation tag Z.

The lack of other categories is indeed debatable – e.g. for the noun-like abbreviations it is reasonable to expect them to have cases, numbers and genders. Our decision was based on the resulting simplicity and the evasion of the need to disambiguate uninflected abbreviations for the values.

Reflexive morphemes *sa* and *si* are treated in a special way. They can be a part of a verb as a reflexive morpheme; however they are detached from the verb itself and even their position in the sentence can somewhat vary. The situation is complicated by the fact that *sa* and *si* are also (reflexive) pronouns – an abbreviated form of *seba* and *sebe*, and the distinction of a verbal morpheme and a pronoun is very subtle. If there are more of the pronouns/morphemes in the clause, they usually fuse into one, e.g in the sentence *bojím* **si** *priznať pravdu* there are two verbs *bojím* **sa** and *priznať* **si**.

We solved the problem by assigning a special tag R to both *sa* and *si*, regardless of their function. Unrelated uses of *si* as a 2nd person singular of the verb *byť* and the use of *sa* as a (poetic) particle in (fixed) expressions *sem sa*, *hor sa* are marked as a verb and a particle, respectively.

Interjections are tagged with J. Accordingly with the traditional grammars, we also mark greetings as interjections (*ahoj/*J, *ahojte/*J, *čau/*J, *čaute/*J …), where the lemma is always identical with the word form.

Numbers written as digits are marked with the tag 0 (the digit zero). Both Arabic and Roman numerals are recognised, the lemma is identical to the word form (except for misspellings, where the lemma is normalised to the 'correct' form, e.g. *l984/*0 with a leading letter instead of digit will be lemmatised as *1984*).

Undefined part of speech (residual) is a token that cannot have its part of speech determined – the reason is usually that it is a part of a multiword expression that has been tokenised as several separate tokens. It is tagged with the symbol `Q`. Examples include hyphenated compounds (*sociálno/*`Q` *-/*`Z` *ekonomický/*`AAis1`), components of foreign proper names (*New/*`Q` *York/*`SSis4:r`), but also tokens that are not considered standalone words by traditional grammar theories – expressions like *po/*`Q` *anglicky/*`Dx`, *na/*`Q` *modro/*`Dx`, where the whole expression will be considered an adverb.

Foreign language citation is reserved for citation elements, i.e. foreign language words that appear to be foreign elements in the text (neither loanwords, nor commonly used proper names). Typically, these are short citations, names of books, movies etc. The symbol for this tag is `%` `U+0025 PERCENT SIGN`.

Non-word element is anything that is neither a word nor punctuation. Typically, these are remnants of incorrect conversion (which would not be there in an ideal world), (pseudo)graphical elements, fancy paragraph separators. A simple test deployed in the tagging process is to consider non-word elements tokens that do not belong to a fixed set of common punctuation characters and that do not consist of alphanumeric characters. The symbol for a non-word element is `#` `U+0023 NUMBER SIGN`. A typical example is the copyright sign ©*/*`#`.

7 CONCLUSION

We have described the tagset designed and used in the Slovak National Corpus. The tagset is used in a morphological database of Slovak words and in the manually annotated corpus of Slovak language, *r-mak*. The database and the manually annotated corpus are then used to train an automatic morphological tagger *morče* (Votrubec, 2006) developed at the Faculty of Mathematics and Physics, Charles University and used originally to tag the Czech language. Morče is used for automatic lemmatisation and tagging of the whole Slovak National Corpus and other Slovak language corpora and subcorpora. The tagset has become de facto the standard tagset used in automatic morphosyntactic analysis and tagging of Slovak language texts. The complete tagset tables with examples can be found online at `http://korpus.sk/morpho.html`.

REFERENCES

Vladimír BENKO, Jana HAŠANOVÁ, and Eduard KOSTOLANSKÝ (1998), *Model morfologickej databázy slovenčiny. Počítačové spracovanie jazyka*, Pedagogická fakulta Univerzity Komenského, Bratislava, Slovakia.

Roger COMTET (1997), *Grammaire du russe contemporain*, Presses Universitaires du Mirail.

Łukasz DĘBOWSKI (2001), Tagowanie i dezambiguacja, in *Prace IPI PAN 934*, Instytut Podstaw Informatyki PAN, Warsaw, Poland.

Ludmila DIMITROVA, Tomaž ERJAVEC, Nancy IDE, Heiki Jaan KAALEP, Vladimir PETKEVIČ, and Dan TUFIŞ (1998), Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages, in *Proceedings of the COLING-ACL'98*, pp. 315–319, Montréal, Québec, Canada.

Ladislav DVONČ, Gejza HORÁK, František MIKO, Jozef MISTRÍK, Ján ORAVEC, Jozef RUŽIČKA, and Milan URBANČOK (1966), *Morfológia slovenského jazyka*, Vydavateľstvo Slovenskej akadémie vied, Bratislava, Slovakia.

Sašo DŽEROSKI, Tomaž ERJAVEC, and Jakub ZAVREL (2000), Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagset, in *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp. 1099–1044, ELRA, Paris, France.

Radovan GARABÍK (2006), Slovak morphology analyzer based on Levenshtein edit operations, in *Proceedings of the WIKT'06 conference*, pp. 2–5, Institute of Informatics SAS, Bratislava, Slovakia.

Radovan GARABÍK (2011), Slovak MULTEXT-East Morphology tagset, *Jazykovedný časopis*, (1):19–39.

Radovan GARABÍK, Lucia GIANITSOVÁ, Alexander HORÁK, and Mária ŠIMKOVÁ (2004), Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu, URL `http://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf`, Internal documentation.

Radovan GARABÍK, Daniela MAJCHRÁKOVÁ, and Ludmila DIMITROVA (2009), Comparing Bulgarian and Slovak Multext-East morphology tagset, in *Organization and Development of Digital Lexical Resources*, pp. 38–46, Dovira Publishing House, Kyiv, Ukraine.

Jan HAJIČ (2004), *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Karolinum, Charles Univeristy Press, Prague, Czech Republic.

Jan HAJIČ (2000), Morphological Tagging: Data vs. Dictionaries, in *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pp. 94–101.

Jan HAJIČ and Barbora VIDOVÁ-HLADKÁ (1997), Morfologické značkování korpusu českých textů stochastickou metodou, 4(58):288–304.

Matej POVAŽAJ, editor (2003), *Krátky slovník slovenského jazyka. 4., doplnené a upravené vydanie*, Veda, Bratislava, Slovakia.

Emil PÁLEŠ (1994), *SAPFO. Parafrázovač slovenčiny. Počítačový nástroj na modelovanie v jazykovede*, Veda, Bratislava, Slovakia.

Radek SEDLÁČEK (2001), A new Czech morphological analyser ajka, in *Proceedings of the TSD, Czech Republic*, pp. 100–107, Springer Verlag.

Serge SHAROFF, Mikhail KOPOTEV, Tomaz ERJAVEC, Anna FELDMAN, and Dagmar DIVJAK (2008), Designing and Evaluating a Russian Tagset, in Nicoletta CALZOLARI, Khalid CHOUKRI, Bente MAEGAARD, Joseph MARIANI, Jan ODJIK, Stelios PIPERIDIS, and Daniel TAPIAS, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, URL `http://www.lrec-conf.org/proceedings/lrec2008/`.

Pavel ŠMERK (2010), A New Data Format for Czech Morphological Analysis, in *Proccedings of Recent Advances in Slavonic Natural Language Processing*, pp. 3–8, Tribun EU, Karlova Studánka, Czech Republic, URL `http://www.fi.muni.cz/sojka/download/raslan2010/raslan10.pdf`.

Jan VOTRUBEC (2006), Morphological Tagging Based on Averaged Perceptron, in *WDS'06 Proceedings of Contributed Papers*, pp. 191–195, Matfyzpress, Charles University, Praha, Czech Republic.

The Bulgarian National Corpus: Theory and Practice in Corpus Design

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova
Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria

ABSTRACT

The paper discusses several key concepts related to the development of corpora and reconsiders them in light of recent developments in NLP. On the basis of an overview of present-day corpora, we conclude that the dominant practices of corpus design do not utilise the technologies adequately and, as a result, fail to meet the demands of corpus linguistics, computational lexicology and computational linguistics alike.

We proceed to lay out a data-driven approach to corpus design, which integrates the best practices of traditional corpus linguistics with the potential of the latest technologies allowing fast collection, automatic metadata description and annotation of large amounts of data. Thus, the gist of the approach we propose is that corpus design should be centred on amassing large amounts of mono- and multilingual texts and on providing them with a detailed metadata description and high-quality multi-level annotation.

We go on to illustrate this concept with a description of the compilation, structuring, documentation, and annotation of the Bulgarian National Corpus (BulNC). At present it consists of a Bulgarian part of 979.6 million words, constituting the corpus kernel, and 33 Bulgarian-X language corpora, totalling 972.3 million words, 1.95 billion words (1.95×10^9) altogether. The BulNC is supplied with a comprehensive metadata description, which allows us to organise the texts according to different principles. The Bulgarian part of the BulNC is automatically processed (tokenised and sentence split) and annotated

Keywords: corpus design, Bulgarian National Corpus, computational linguistics

at several levels: morphosyntactic tagging, lemmatisation, word-sense annotation, annotation of noun phrases and named entities. Some levels of annotation are also applied to the Bulgarian-English parallel corpus with the prospect of expanding multilingual annotation both in terms of linguistic levels and the number of languages for which it is available. We conclude with a brief evaluation of the quality of the corpus and an outline of its applications in NLP and linguistic research.

1 INTRODUCTION

Since the first structured electronic corpus, the Brown Corpus (Francis and Kučera, 1964), corpora have been increasingly used as a source of authentic linguistic data for theoretical and applied research. Corpus-based studies have been employed in various areas of linguistics, such as lexicology, lexicography, grammar, stylistics, sociolinguistics, as well as in diachronic and contrastive studies (Meyer, 2002).

Traditional definitions of a corpus emphasise different aspects. A corpus is typically viewed as a collection of authentic linguistic data that may be used in linguistic research (Garside *et al.*, 1997). Sinclair (2005) adds to this definition the storage format and the selection criteria: "*A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.*" Finally, annotation at different linguistic levels (phonological, lexical, morphological, morphosyntactic, syntactic, semantic, discourse and stylistic) amplifies the corpus's value by extending its functionalities and applications (McEnery *et al.*, 2006). One of many different definitions states: *A corpus is a large collection of language samples, suitable for computer processing and selected according to specific (linguistic) criteria, so that it represents an adequate language model.* (Koeva, 2010).

With the increased development of language technologies, the applications of corpora have been extended to all areas of computational linguistics and natural language processing (NLP). Corpora have become an indispensable resource for generating training sets for machine learning, language modelling, and machine translation. These developments have led to the necessity for reconsidering the traditional notions in corpus linguistics. As a result, we propose a corpus design based on automatic collection of very large monolingual and

multilingual (and in particular parallel) corpora that cover a wide variety of styles, thematic domains, and genres.

This paper contributes to the discussion on the perspectives of corpus development in three ways: (i) by reconsidering several key traditional principles underlying corpus design, (ii) by proposing an approach in corpus design based on the revision of those fundamentals in light of recent advances in NLP technologies, (iii) by illustrating how the proposed model is applied in the Bulgarian National Corpus (BulNC).

The study is placed in the context of well-known corpora, both mono- and multilingual (Section 2), with an outline of their general features. The concepts of corpus size, balance, and representativeness are discussed in Section 3. In the same section we present our concept of corpora, which integrates the best practices of traditional corpus linguistics with the potential of the latest technologies for web crawling and language processing. Section 4 presents the process of compiling, structuring, documenting, and annotating the BulNC, followed by a brief evaluation of the quality of the corpus and an outline of some current applications.

2 OVERVIEW OF CONTEMPORARY MONOLINGUAL AND MULTILINGUAL CORPORA

The last decades have seen the compilation of large mono- and multilingual corpora for a lot of languages, including some less-resourced ones, Bulgarian among them. The brief overview illustrates the current standards in corpus design and compilation and provides a point of departure for comparison with the proposed paradigm.

2.1 *Large monolingual corpora*

1. At the time of its creation, the British National Corpus[1] (BNC) was one of the biggest (100 million words) existing corpora. Being compiled according to carefully devised principles and classification criteria[2], it set the standards for general monolingual synchronic corpora for quite some time. The BNC represents not

[1] `http://www.natcorp.ox.ac.uk`

[2] `http://www.natcorp.ox.ac.uk/corpus/creating.xml`

only written, but also spoken language, respectively 90% and 10% of the samples. It is POS-tagged, lemmatised, and supplied with detailed metatextual information. The corpus (text and annotated data) can be searched both online – through various search tools, and offline using XAIRA[3].

2. The Corpus of Contemporary American English[4] (COCA) is a 450+ million-word corpus currently in progress with an increase rate of 20 million words per year. The texts are evenly divided between 5 categories – spoken language, fiction, popular magazines, newspapers, and academic writing (Davies, 2010), each category currently containing 90 to 95 million tokens (as of June 2012). The corpus provides a web search interface (shared with the Google Books corpora) that allows searches for regular expressions and specifications for POS, lemma, collocations, frequency and distribution of synonyms. The queries may be refined in terms of genre or time period.
3. The Slovak National Corpus[5] (SNK) contains more than 719 million tokens[6]. The texts are divided into several categories with the following distribution: journalism (73%), literary texts (14%), professional texts (12%), and other (1%). A subcorpus of 1.2 million tokens, manually annotated with morphological tags, has also been compiled. The SNK and its subcorpora can be searched with a CQL (Corpus Query Language) compatible query syntax (Christ and Schulze, 1994) through a web interface or via the Bonito client[7], cf. the Czech National Corpus.
4. The Croatian National Corpus[8] (HNK) includes about 101 million words of mainly contemporary Croatian texts that cover different media, genres, styles, fields, and topics. They fall into the categories of informative texts (74%), fiction (23%), and mixed texts (3%), with further subdivision within these categories. The morphological tagset used in the HNK annotation is Multext-East-

[3] `http://www.natcorp.ox.ac.uk/tools/index.xml`
[4] `http://corpus.byu.edu/coca/`
[5] `http://korpus.sk`
[6] The version released at the beginning of 2011
[7] `http://korpus.juls.savba.sk/usage_en.html`
[8] `http://www.hnk.ffzg.hr/cnc.htm`

compatible, and the corpus can be searched offline through the Manatee/Bonito server-client[9].

5. The Russian National Corpus[10] (RNC) comprises more than 300 million words of texts ranging from the middle of the 18th century to the present day. The main part of the corpus, about 100 million words, consists of contemporary texts of three general categories: fiction (40%), non-fiction (56%) and recordings of public and spontaneous speech (4%), with a detailed internal classification[11]. The corpus has been automatically supplied with morphosyntactic annotation, and parts of it have been manually verified and disambiguated. A portion of the corpus has also been annotated with syntactic dependencies and semantic roles. Lexical-semantic information, covering taxonomic, mereological, topological and other features of words, has been assigned. The RNC provides a web interface for detailed search in the whole corpus and its subcorpora for words and phrases, grammatical (POS, morphology), syntactic, and semantic (taxonomy, evaluation and mereology) features.

6. The Czech National Corpus[12] (CNC) (Kocek *et al.*, 2000) was started in the 1990s. Since then it has been constantly growing and according to the latest published estimates currently amounts to 1.3 billion words. It consists of a number of subcorpora, among them several balanced subcorpora of 100 million words each, compiled every several years, the latest version being SYN2010. The distribution of texts across categories is as follows: fiction (40%), technical literature (27%) and journalism (33%), with more elaborate subdivision within these categories. Most of the written corpora in the CNC are annotated. The CNC can be searched for words and phrases using exact match and regular expressions both online and offline through the Corpus manager Manatee and the client Bonito[13].

[9] `http://www.hnk.ffzg.hr/pretraga_en.html`
[10] `http://www.ruscorpora.ru/`
[11] `http://www.ruscorpora.ru/en/corpora-stat.html`
[12] `http://ucnk.ff.cuni.cz`
[13] `http://www.textforge.cz/download`

7. The National Corpus of Polish[14] (NCP) (Bański and Przepiórkowski, 2010) contains fiction, daily newspapers, specialised periodicals and journals, transcripts of conversations and Internet texts, amounting to approximately 1 billion words. A balanced 250-million-word subcorpus extracted from the NCP has been compiled (Przepiórkowski, 2011), and part of the NCP, approximately 1.2 million words, was manually annotated. The corpus can be searched online through two search engines (Poliqarp and PELCRA) that allow queries for words and regular expressions; the former also searches for morphological tags and the latter offers collocation extraction. The search may be further refined with editorial and descriptive (genre or domain) metadata.
8. The German Reference Corpus[15] (DeReKo) amounts to 5.4 billion words (Bański *et al.*, 2012).The concept of the corpus explicitly rejects the feasibility of balance and representativeness (Kupietz *et al.*, 2010). The corpus is conceived as a versatile "primordial" sample from which specialised subsamples, or "virtual" corpora, are drawn. The development of the corpus is focused on the maximisation of size and stratification, rather than on the composition of specialised subsamples. The corpus includes POS annotation, partial morphological disambiguation, named entities, and syntactic dependencies. The corpus can be searched offline through the COSMAS II client[16].
9. A number of large corpora have recently come into existence, with size ranging from several (Baroni and Kilgarriff, 2006; Pomikálek *et al.*, 2009), through dozens (Pomikálek *et al.*, 2012), to hundreds of billions of words (Google Books Corpora, GBC[17], the largest being the 200-billion-word GBC of American English). What distinguishes these from the rest of the discussed corpora is that they represent a different type of approach to corpus creation, since they are collected fully automatically from web content. The GBC web search interface allows queries according to several criteria:

[14] `http://nkjp.pl`

[15] the Archive of General Reference Corpora of Contemporary Written German, `http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html`

[16] `http://www.ids-mannheim.de/cosmas2/`

[17] `http://googlebooks.byu.edu/`

exact words or phrases, regular expressions, POS, lemma, collocations, frequency and distribution of synonyms, with further refinement in terms of genre or time period.

This brief outline shows that the dominant and constant tendency is for corpora to aim at a size ranging from several hundred million up to over a billion words.

All corpora are at least partly annotated and provide online or offline search interfaces. The differences lie in the quantity of the annotated data and the levels of annotation. The minimum annotation is generally POS tagging. Many corpora also include morphosyntactic annotation and lemmatisation, and some provide syntactic (e.g., DeReKo) or semantic annotation (e.g., COCA, GBC).

Some of the corpora discussed here follow predefined structure and classifications, whereas others abandon balance and representativeness in favour of size. The design criteria differ not only when it comes to coverage and distinction of textual categories and subcategories, but, more fundamentally, in the underlying assumptions. Balance is viewed as the equal representation of predefined text categories (Davies, 2010), or as a distribution of texts proportional to language production (Atkins *et al.*, 1991) or language reception estimated according to various criteria. Some authors, involved in the compilation of the previously discussed corpora, have proposed assessments of language reception on the basis of stylistic (Przepiórkowski *et al.*, 2010), sociological (Čermak and Schmiedtová, 2003), and marketing (Tadić, 2002) surveys.

The following trends emerge with respect to the relationship between size, balance and representativeness:

- Creation of corpora according to a predefined methodology that is considered sufficiently adequate to ensure corpus balance and representativeness (1-5).
- Development of large unbalanced corpora paired with static balanced subcorpora that are compiled in accordance with a carefully devised structure (6-7).
- Compilation of large unbalanced corpora that enables the extraction of subcorpora based on metadata description (8).

- Compilation of very large unbalanced corpora from the web whose structure and content are not concerned with balance and representativeness (9).

2.2 *Large parallel corpora*

1. Some of the major parallel corpora that have been largely drawn on by the NLP community are multilingual repositories of publicly available legal, administrative or journalistic texts, such as: the European Parliament Proceedings Parallel Corpus[18] (Europarl), the Canadian Hansard Corpus of parliamentary proceedings; the News Commentary Corpus[19]; the JRC-Acquis Multilingual Parallel Corpus[20] of legal texts, the EU Official Journal[21], MultiUN[22]. The OPUS collection[23] includes a set of various corpora – administrative (e.g., the EMEA corpus of administrative medical texts), news (including the SETimes corpus of news in eight Balkan languages and English), etc.
   These corpora are distinguished from traditional ones in that the data have been compiled for a different purpose and have subsequently been employed as corpora. Therefore not all of them are annotated. OPUS, EuroParl, and JRC-Acquis are tokenised, sentence-segmented and sentence-aligned. Parts of the OPUS collection are POS-tagged for some languages, with word alignment currently under way and dependency parsing envisaged in the near future. A part of the Hansard corpus is also sentence-split and aligned.
2. The Czech-English parallel corpus (CzEng) comprises 206.4 million tokens in Czech and 232.7 million tokens in English, distributed across 7 source domains: fiction, EU legislation, movie subtitles, parallel webpages, technical documentation, news, and texts from Project Navajo[24]. The predominant domains are fiction and legislation. The texts have gone through automatic sentence-

[18] http://www.statmt.org/europarl
[19] http://www.statmt.org/wmt11/translation-task.html
[20] http://langtech.jrc.it/JRC-Acquis.html
[21] http://eur-lex.europa.eu/en/index.htm
[22] http://www.euromatrixplus.net/multi-un/
[23] http://opus.lingfil.uu.se/
[24] http://ufal.mff.cuni.cz/czeng/czeng10/

splitting and alignment. Morphosyntactic tagging, lemmatisation, word alignment, surface and deep-level syntactic annotation are provided (Bojar *et al.*, 2012).

3. The Hunglish corpus[25] is a sentence-aligned parallel corpus of Hungarian and English containing 34.6 million Hungarian and 44.6 million English words. The texts cover a number of varied domains: literature, religion, international law, movie subtitles, software documentation, magazines, and business reports (Varga *et al.*, 2005). The corpus has been tokenised with the rule-based HunToken tokeniser and stemmed with the Hunmorph morphological analyser.
4. The Polish-Russian Parallel Corpus[26] consists of 50 million words equally divided between Polish originals with their Russian translations and the other way round. The corpus includes classical and modern literature as well as legal and journalistic texts. The texts are annotated according to the annotation schemes of the National Corpus of Polish and the Russian National Corpus.
5. The Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles[27] is a corpus of 440,000 parallel sentences made up of 12 million Japanese words and 11.5 million English words. The texts concerning Kyoto and other specific topics, such as traditional Japanese culture, religion, and history, are manually translated into English, aligned and verified. The corpus was used for the development and evaluation of Japanese-English machine translation systems in the Kyoto Free Translation Task (Neubig, 2011).

Many of the available parallel corpora are of modest size, especially in comparison with monolingual corpora, and as a rule they belong to a limited number of domains determined by the availability of parallel texts. For the most part these corpora are compiled automatically by web crawling or by downloading publicly available parallel collections. More varied content can be obtained from publishers or through manual compilation, but these methods are less efficient. A third source of parallel data has been the translation of monolingual corpora; Xu and Sun (2011), among others, have experimented with

[25] `http://mokk.bme.hu/resources/hunglishcorpus/`

[26] `http://www.pol-ros.polon.uw.edu.pl/`

[27] `http://alaginrc.nict.go.jp/WikiCorpus/index_E.html`

machine translation for increasing parallel data for less common languages.

Due to the limited domain and genre diversity of parallel texts, balance and representativeness are usually considered irrelevant. Three dominant patterns of parallel corpus design emerge:

- Acquisition of corpora from repositories of parallel content (1). Being publicly available, these collections are often reused in other corpora either as raw text or with the supplied annotation.
- Compilation (preferably automatic) of parallel corpora that aims at reflecting the diversity of monolingual corpora, possibly using readily available corpora (2-4).
- Construction of parallel corpora by means of human translation (5) or machine translation of the original content.

To conclude, there is considerable heterogeneity among the existing monolingual and parallel corpora in terms of size, design criteria, annotation principles, etc. At the same time, neither the possibilities of the modern technologies, nor the enormous amount of available data are used rationally enough to serve the needs of NLP. Moreover, traditional services for extraction of concordances and collocations fail to meet the needs of modern corpus linguistics and computational lexicography. The automatic collocation dictionaries extracted through "sketch grammars" and an algorithm for finding "good dictionary examples" allow a more efficient access to corpus data (Kilgarriff *et al.*, 2009).

2.3 *An overview of Bulgarian corpora*

The work on corpora for Bulgarian began in the 1990s with the compilation of relatively small text collections for specific purposes.

1. Two corpora of Spoken Bulgarian have been created in the 1990s[28] (Nikolova, 1987; Aleksova, 2000).
2. Further efforts were focused on building large reference corpora such as the BulTreeBank Text Archive (since 2000) with 15% of the texts coming from fiction, 78% from newspapers and about 7% excerpted from legal and government texts and others (Simov

[28] `http://folk.uio.no/kjetilrh/bulg/Aleksova/index.html` and `http://folk.uio.no/kjetilrh/bulg/Nikolova/`

et al., 2002). This corpus recently evolved into the Bulgarian National Reference Corpus[29]. The corpus interface executes queries allowing exact matches or regular expressions.

3. The "Brown" Corpus of Bulgarian[30] (BCB) (Koeva *et al.*, 2006), compiled in the period 2001 to 2005 as a general corpus of contemporary Bulgarian, is one of the Bulgarian corpora that closely follow a clearly established methodology, namely that of the Brown Corpus of Standard American English (Francis and Kučera, 1964). Text samples can be searched using queries for exact matches or regular expressions. The "Brown" Corpus of Bulgarian with full-length texts[31] (FullBrown), consisting of the originals from which the BCB 2000-word excerpts were sampled, is included as an integral part of the Bulgarian National Corpus and may be searched through its web interface.
   Concomitantly, a number of Bulgarian annotated corpora have been developed, covering POS tagging, word sense annotation, annotation of dependency structure, and sentence and clause alignment.
4. The Bulgarian POS-Tagged Corpus (BulPosCor) totals 174,697 words, each of them manually annotated with the context-relevant POS and morphosyntactic features.
5. The Bulgarian Sense-Annotated Corpus (BulSemCor) amounts to 95,119 lexical items, covering both single words and multiword expressions. Each of them has been POS-tagged, lemmatised, and assigned a synonym set from the Bulgarian Wordnet (Koeva *et al.*, 2006) that best corresponds to the sense of the lexical item in the particular context.
   Both BulPosCor and BulSemCor are integrated into the BulNC and can be accessed through its web search interface (see Section 4.6.1), as well as through specially developed interfaces[32].
6. The BulTreeBank is a syntactically annotated corpus developed within the HPSG framework (Simov and Osenova, 2004). Access to the data may be gained by submitting a contact form. The

[29] `http://www.webclark.org`
[30] `http://dcl.bas.bg/Corpus/home_en.html`
[31] `http://dcl.bas.bg/en/corpora_en.html`
[32] `http://dcl.bas.bg/poscor/en/`, `http://dcl.bas.bg/semcor/en/`

HPSG-based annotation has later been converted into syntactic dependency annotation covering 214,000 tokens, or slightly more than 15,000 sentences (Chanev *et al.*, 2006).

7. The Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC) (Koeva *et al.*, 2012a), a parallel sample from the Bulgarian National Corpus, comprises 176,397 tokens for Bulgarian and 190,468 for English. It is supplied with syntactic annotation for sentence and clause boundaries, relations between syntactically linked clauses (i.e., coordination or subordination) and clause-introducing words and phrases.

The main purpose of corpora 4 to 7 is to serve as training and test corpora in the development of various automatic annotation tools for multi-level annotation with sufficient accuracy and coverage. Most of the parallel corpora involving Bulgarian are purpose-driven and cover specific domains, such as administrative texts or fiction, which are widely available in parallel versions and hence easily collected.

The Bulgarian subcorpus in the JRC-Acquis corpus of EU legislation documents contains 16.1 million tokens (Steinberger *et al.*, 2006). The Bulgarian part of the SEE-ERA.NET Administrative Corpus (SEnAC) consists of excerpts from the Acquis communautaire, about 1.5 million tokens and 60,389 translation units, each containing one sentence translated into 8 languages (Tufiş *et al.*, 2009). The EuroParl Corpus of proceedings of the European Parliament includes approximately 6 million tokens in Bulgarian (Koehn, 2005).

The Multext-East corpus incorporates the original and translations into six languages of George Orwell's novel *Nineteen Eighty-Four*, with the Bulgarian part amounting to 54,823 tokens (Dimitrova *et al.*, 1998). In the SEE-ERA.net Fiction Corpus (SEnFC), consisting of translations of Jules Verne's novel *Around the World in 80 Days* into sixteen languages, the Bulgarian part adds up to 58,678 tokens (Tufiş *et al.*, 2009). The Cultural Greek-Bulgarian Corpus is a bilingual collection of literary and folklore texts containing approximately 350,000 tokens (Giouli *et al.*, 2009). The extended RuN-Euro Corpus includes a small Bulgarian part of 366,329 tokens (Grønn and Marijanovic, 2010), and ParaSol (von Waldenfels, 2006, 2011), a corpus of fiction texts, includes a Bulgarian subcorpus of over 2 million tokens as of June 2011. The Bulgarian-Polish-Lithuanian Parallel Corpus (Dim-

itrova *et al.*, 2009) contains more than one million words and combines texts from more than one domain – fiction texts and administrative texts (EU documents). Some parallel corpora in the OPUS collection (Tiedemann, 2009) include Bulgarian – medical documents by the European Medicines Agency, movie subtitles, and the SETimes news corpus.

Smaller and purpose-driven corpora, such as the *Nineteen Eighty-Four* (Multext-East) and the SEnAC and SEnFC corpora (SEE-ERA.NET), are tokenised, lemmatised, POS-tagged and aligned at sentence level. Annotation of larger corpora is usually limited to tokenisation, sentence splitting, and alignment, e.g. the OPUS collection, with the tendency to be extended to other levels of annotation.

Due to the limited amount of translations between particular pairs of languages, the interest in comparable corpora has been growing in the last decades. Still, only a small number of comparable corpora involve Bulgarian. The Multext-East comparable corpora with subcorpora of Bulgarian, Czech, English, Estonian, Hungarian, Romanian, and Slovene, include fiction and newspaper data (Dimitrova *et al.*, 1998). The Bulgarian-Croatian Comparable Corpus (Bekavac *et al.*, 2004) contains newswire texts, 393,000 tokens for Bulgarian and 1.3 million for Croatian. The Bulgarian-Polish-Lithuanian Comparable Corpus comprises fiction and electronic media documents balanced in size across the three languages (Dimitrova *et al.*, 2009).

This overview suggests that the existing Bulgarian corpora share most of the merits of the corpora compiled for other languages, and suffer from similar shortcomings, further aggravated by their smaller size and limited diversity, as well as the restricted availability of both monolingual and parallel data. On the positive side, most of the manually annotated corpora conform to the best annotation practices and have been employed in the development of various NLP applications for Bulgarian.

3 KEY FEATURES OF CORPORA

Apart from their use in traditional corpus linguistics and computational lexicography, contemporary corpora have been increasingly employed in developing language models, translation models and in training machine learning algorithms. However, despite the rapid de-

velopment of technologies and the vast amount of electronic data, the available corpora largely adhere to long-established tradition: they aspire to represent a balanced sample of language and for that reason constitute collections of carefully selected, often fixed-size, text excerpts. They are being explored with outdated methods and tools, which limits their use to extraction of concordances and collocations.

In the context of the dynamically evolving web and with more and more mono- and multilingual corpora becoming available, the traditional understanding of corpus design has been undergoing reconsideration. Some of the most prominent corpus features: size, balance, and representativeness (Xiao, 2010) will be discussed in the following subsections. We then proceed to propose a general approach for corpus development based on *automatic compiling*, *detailed metadata description*, and *multiple annotation*. This, we believe, will result in dynamic enlargement and efficient management of corpus data.

3.1 *Corpus size revisited*

A corpus large enough for empirical studies on language might not contain sufficient occurrences of specific and rare language phenomena for drawing statistically valid conclusions (Banko and Brill, 2001; Keller and Lapata, 2003; Kilgarriff and Grefenstette, 2003). Consequently, for building probabilistic models, larger amounts of data are needed, as large data, even if they are noisy, yield more reliable models than estimates based on smaller, limited datasets. After exploring the performance of a number of machine-learning algorithms for disambiguation when the size of the training corpus was increased from a million to a billion words, Banko and Brill (2001) concluded that the performance of any algorithm improves with data size, although the optimal data size varies with different algorithms[33]. This assertion is reflected in the rationale behind the web-as-a-corpus framework (Kilgarriff and Grefenstette, 2003), where the case is made for the necessity of making vast amounts of Internet texts available for processing and querying. Scientists have long since realised that the largest corpus is the web and that what primarily keeps Internet data

[33] An important insight made by Curran and Osborne (2002) in criticising Banko and Brill (2001) is that the benefits of large amounts of data are better experienced when size is combined with sophisticated statistical language models.

from becoming a real corpus is the lack of linguistically focused metadata and annotation.

The major size-related concern for corpus linguistics is how to define optimality – in terms of corpus size and in terms of sample size. The criterion for optimal corpus size aims at ensuring adequate coverage of lexical diversity, estimated with respect to wordstock, thematic domains, genres or language phenomena, while the criterion for text sample size takes into account the balance between texts, as well as their diversity. Several attempts at approximating an "ideal" size and structure, defined intuitively or empirically, have been made. Yang *et al.* (2000) try to estimate a corpus size that would be sufficient for obtaining the core vocabulary, while Chevelu *et al.* (2007) propose an algorithm for calculating an optimal corpus design that ensures coverage of a preset description of phonological attributes. What qualifies as optimal corpus size is still an open question, contingent on the particular linguistic or lexicographic task.

We shall illustrate the relation between corpus size and lexical diversity by a comparison between the "Brown" corpus of Bulgarian and FullBrown. The former consists of around one million words in 500 fixed-size samples of approximately 2000 words with adjustment to sentence boundaries; the latter includes the full-length originals of the BCB excerpts and totals 4.5 million words. The Bulgarian "Brown" corpus has 112,130 unique tokens, of which 61,162 (6.12% of all corpus tokens) appear only once. FullBrown contains 256,413 unique tokens and 130,230 tokens (2.89% of all corpus tokens) have a frequency of 1.

The early corpus tradition used a fixed size of the text samples to ensure balance and diversity of data and to avoid the skewing that might result from including large texts. Although limiting the size of samples is still appropriate for balanced and domain- and purpose-specific corpora, the above example shows that the inclusion of full texts contributes to language diversity and helps overcoming data sparsity.

The approach we adopt is based on two assumptions: that larger corpora are better suited to language analysis, irrespective of the particular task; and, that these resources, if properly documented and annotated, may also serve as a reliable source from which smaller, uniformly processed, different-sized subcorpora can be extracted, thus

eliminating the need for ad-hoc building of standalone fixed-structure corpora. Therefore we include the full versions of the texts in the corpus, as this allows us to extract comprehensive statistical metadata for the number of tokens, words, lemmas, clauses, sentences, and specific grammatical constructions that would enable further extraction of subcorpora, such as the one compiled for the development of the Bulgarian Sense-Annotated Corpus (BulSemCor; cf. Koeva *et al.* 2011).

3.2 *Balance and representativeness reconsidered*

The size of contemporary corpora comes at the expense of their structure, as they are usually created by collecting vast amounts of data at a fast rate. The predefined design criteria that used to be the organising principle of post-Brown corpora turn out to be empirically refuted and in need to be redefined in terms of the availability of various text types.

Representativeness is associated with the adequate coverage of the varieties of language use, while balance concerns the linguistically relevant distribution of texts across categories (Sinclair, 2005). Although these corpus features have been the focus of extensive study (Leech, 1991; Atkins, 1992; Biber, 1993; Sinclair, 2005; McEnery *et al.*, 2006), definitive qualitative or quantitative criteria for ensuring or evaluating them have not been convincingly established.

As a consequence, in traditional accounts, where the notions of balance and representativeness are defined in terms of supposedly relevant linguistic coverage and proportions of total language production, they remain tentative notions (Manning and Schutze, 1999; Kilgarriff and Grefenstette, 2003; Kupietz *et al.*, 2010). Moreover, defining them in this fashion is not adequate when it comes to the requirements posed to corpora by NLP. The new demands have called for a shift from compiling corpora that are carefully proportioned in terms of sample size and text types to expanding the quantity of the data.

Below, we attempt to redefine the relationship between size, balance, and representativeness in a data-driven perspective.

Representativeness is recast in terms of the range and diversity of text categories accompanied by enrichment of the sampling methodology. Since balance is hard to maintain for dynamic (constantly evolving) corpora, we suggest that instead of trying to maintain it for the

whole corpus, we extract different balanced subcorpora based on a large set of criteria (both preset or user-defined) such as time period, thematic domain, genre, author, density or distribution of certain language phenomena, etc.

We focus on amassing large amounts of texts that cover a variety of languages, media type, styles, domains, genres, and topics. The dynamic enlargement of the corpus, including the growth rate, the range of samples, and their quantity, is determined by the availability of texts on the web rather than by a preset model. Corpus structure is ensured through detailed metadata organised in a comprehensive classification of categories. The detailed metadata description allows for easy compilation of general, domain- and purpose-specific subcorpora with a fixed structure or predefined features. The metadata classification scheme is flexible, in order to match the new texts that are constantly being included in the corpus.

3.3 *Extended metadata and linguistic annotation*

Metadata describe the properties of the text samples in the corpus and are external to the text itself. Burnard (2005) emphasises the importance of metadata and the need for them to be as detailed as possible so that one may be able to determine the relevance of a given linguistic resource to one's own purposes. In the proposed framework, the classification suggested by Burnard (2005) is adopted as a baseline description of the text metadata and the annotation of the texts.

1. Editorial – information about texts in relation to their original source (source, author, date of publishing, etc. Here we include information about language, direction of translation, name of the translator, etc.);
2. Descriptive – classificatory information such as style, domain, and genre;
3. Administrative – documentary information about the texts and the corpus, such as its availability, revision status, etc.;
4. Analytical – various levels of annotation;
5. Statistical – number of tokens, words, general words, domain-specific words, lemmas, noun phrases, phrases, clauses, sentences, etc. In addition to Burnard's classification we include various statistical information.

Extralinguistic metadata about the texts (types 1-3) are built through a combination of automatic and manual techniques with increasing application of the former. Extralinguistic metadata are derived automatically from the HTML markup of the original files or with various heuristics based, for example, on domain-specific or genre-descriptive keywords. Statistical data (type 5) are calculated before or after annotation.

We represent the metadata scheme as a graph (Figure 1) where the nodes are associated with metadata categories and the arcs with binary relations between the nodes, such as *style*, *domain*, and *genre*, etc. For some metadata relations, for instance *style*, the metadata categories are predefined; for others, such as *author*, the categories are an open set. The representation is simplified, e.g. authorship of the text is recorded only once for all kernel and satellite samples in different languages. As a further advantage, graph representation allows flexible extension with new relations and categories and shows where merging or splitting categories is permissible. For example, it is possible to merge the metadata with a database of books' descriptions allowing us to automatically assign publishing dates or obtain translations of the title in different languages. Different "graph mining" algorithms – common subgraph, shortest paths, minimum spanning trees, connectedness, etc. can be used when extracting subcorpora of different types.

Linguistic annotation increases the value of a corpus by making it more *usable*, as various kinds of information may be extracted, more *multifunctional*, as the corpus may be used for different purposes, and more *explicit* with respect to the analysed information (McEnery *et al.*, 2006, p. 30). In our approach, we adopt and supplement the criteria set out by McEnery *et al.* (2006) as follows:

- **Multi-layered** – the more richly annotated the corpus, the broader its range of applications for research and applied studies. Corpus processing needs to cover and accumulate as many levels of linguistic annotation as possible.
- **Compliance with standards** in data formatting and representation of annotation. Unification of various tagsets and data formats, including encodings, is enabled through easy and reliable conversion.

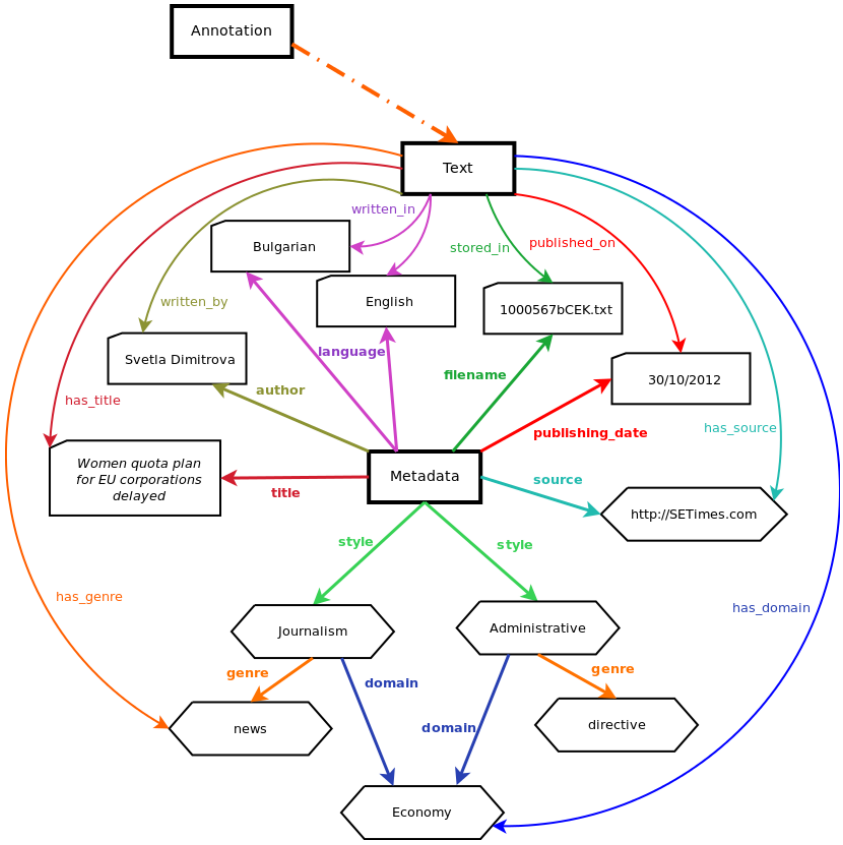


Figure 1: Example of the graph representation of corpus metadata.

- **Uniformity** – a common set of attributes and values for different languages and different media types – text, audio, image, video, and common techniques to manage (accumulate, combine, split, etc.) them. This will facilitate comparative studies and the application of language-independent tools.
- **Consistency** – as annotation of large amounts of texts in most cases is carried out automatically, it is necessary to provide means for validation and evaluation.

The following annotation principles are observed in general, both for manual and automatic annotation (Koeva *et al.*, 2010): the input text remains unchanged; the annotation is performed at consecutive stages and is accumulated as multi-level annotation; the annotation data are represented as attribute-value pairs. Each annotation level

is independent, may be accessed separately and merged with other compatible annotation schemes.

For the abstract representation of annotation an attribute-value formalism is used, in which the attributes are different types of linguistic categories (i.e., word sense, syntactic category, grammatical gender) associated with a set of values, for example *shtastliv* (en: lucky) has the following attributes and values: word sense: *'having or bringing good fortune'*, syntactic category: Adjective, gender: masculine. Ambiguity is not accepted in annotation, so each attribute is assigned a single value. The set of attributes depends on the language features and the granularity of the annotation and is thus open. Binary relations may also be defined between attributes (i.e., common noun – concrete, animate – human), making graph representation possible.

Fully-automated annotation is faster and more consistent although its precision might be lower than manual annotation. Our approach employs primarily automatic annotation, at any level of monolingual and parallel linguistic representation, and to whatever extent possible. Subcorpora with a concentration of a particular grammatical feature (such as singularia tantum) or language construction (such as noun phrases with a prepositional complement) may be extracted on the basis of the annotation.

3.4 *A unified corpus approach*

The need for high-quality monolingual and multilingual corpora further necessitates adjustment of corpus design principles in order to ensure a uniform treatment of monolingual and multilingual corpus parts, with all texts being compiled, documented, processed and accessed within a common framework.

The main source for corpus compilation is and will be the *Internet*, through downloading of readily available text collections or by web crawling. Modern corpus development has to be based on *automatic (and semiautomatic) collection, documentation and annotation* of monolingual and multilingual corpora, while manual work is mainly reduced to defining metadata and annotation schemes, annotation tagsets and the development of gold-standard corpora for training and testing. The requirements for corpus development can be summarized as follows:

1. **Collection** – predominantly automatic compilation of full-text corpus samples by means of web crawling based on preliminary manual and automatic web mining; automatic cleaning of junk formatting, and elimination of duplicates;
2. **Documentation** – detailed metadata extracted from the mark-up of the Internet documents, from the raw data by means of document categorisation and information extraction, and from the annotation by means of statistical processing;
3. **Annotation** – largely automatic linguistic annotation covering different linguistic levels and conforming to uniformity with respect to different languages and different media types.

The corpus design requires a clear-cut structure based on an explicit description of sample categories and explicit mapping between parallel samples in different languages. On the other hand, the corpus structure has to be flexible enough to allow for reorganisation around different categories or languages. This is ensured by a *detailed and consistent metadata documentation* of corpus samples.

Another point of discussion has been whether corpus design should be based on linguistic or extralinguistic criteria (Atkins, 1992; Sinclair, 2005). We subscribe to the idea that text sampling should be based on external criteria derived from the text's communicative function (style, genre, domain, source, year of publication, etc.), rather than on internal criteria that reflect the features of the language of the text (Clear, 1992), since the former afford a more reliable classification, and also to a large extent predetermine the linguistic features of the texts.

The principles for corpus design we have adopted are reflected in the following requirements:

1. Task-independent design ensuring as many monolingual and multilingual data as possible, illustrating different media types with their styles, genres, and domains.
2. Extensibility of the corpus through the inclusion of newly emerging categories attested in language production.
3. Flexibility and robustness of the design in order to facilitate reconsideration and restructuring of classificatory information about the texts. Carefully designed mechanisms for reorganising should

ensure that already included texts are not misclassified after the changes.

4. Adoption of mechanisms for accommodating texts that belong to multiple categories while any additional information is also properly stored and remains accessible.
5. Easy access to the relevant documents, including simple and efficient extraction of information, as well as grouping and regrouping of texts into subcorpora.

This corpus design is proposed in order to maintain simultaneously monolingual and multilingual parallel corpora and allow them to be compiled, preprocessed, annotated, evaluated and accessed through common or compatible tools, compliant with metadata and annotation description schemes, as well as with common (or convertible) annotation tagsets. This approach ensures standardisation, reusability and automation at all stages of corpora development and usage.

Corpus development at the present time needs to take into account the fact that the main purpose of corpora is natural language processing, and should try to answer this field's growing needs of reliable, linguistically enriched multilingual resources. Fulfilling such functions, corpora can successfully serve both corpus linguistics and computational lexicography, as detailed metadata and annotation facilitate the compilation of various domain- and purpose-specific subcorpora.

4 THE BULGARIAN NATIONAL CORPUS

The Bulgarian National Corpus is designed according to the outlined approach. The corpus contains a large variety of texts of different size, media type (written and spoken), style, period (synchronic and diachronic), and languages (Koeva *et al.*, 2012b).

The BulNC started as a monolingual general corpus and has been enlarged constantly, with the latest effort focused on the collection and annotation of parallel data and resulting in the Bulgarian-X Language Parallel Corpus (Bul-X-Cor). The parallel corpora in the BulNC consist entirely of texts that have a Bulgarian counterpart (original or translation) and one or more foreign-language correspondences that can also

be either original or translated. Both the Bulgarian and the foreign versions can be translations from a third language. Bulgarian serves as a pivot language for the parallel corpora, but any X-language is treated equally with respect to text type diversity, preprocessing, metadata scheme, general annotation principles, different levels of annotation, corpus quality evaluation and modes of access for (computational) research and implementations. The corpus may be used for tasks involving any pair of languages available in it. Applying the same principles and methodology used for the Bulgarian part of the BulNC and the Bul-X-Cor ensures, among other things, efficiency in terms of storage, as duplication of files between different parallel corpora is avoided and texts are stored and processed only once, unlike other corpora, such as the corpora in the OPUS collection.

4.1 *Compilation of the BulNC*

Three basic approaches have been applied in the compilation of both the kernel and the satellites:

1. **Using readily available text collections**. The kernel of the Bulgarian National Corpus was first compiled on the basis of the Bulgarian Lexicographic Archive and the Text Archive of Written Bulgarian, which together account for 55.95% of the corpus. Later, two domain-specific corpora from the OPUS collection were included, namely the EMEA corpus (medical administrative texts) and the OpenSubtitles corpus (film subtitles) representing respectively 1.27% and 8.61% of the kernel of the BulNC (see Figure 3). A large amount of news data in the Bulgarian Lexicographic Archive and the Text Archive of Written Bulgarian were provided by the publishers of various Bulgarian newspapers.
   The corpora were either obtained in plain text format or converted to it. Metadata were extracted automatically wherever possible, documented and verified manually in some cases. Full annotation was performed from scratch, even for already annotated texts (OPUS texts are tokenised and sentence-al-igned) to ensure conformity with the adopted principles and annotation standards.
2. **Manual compilation** by browsing the Internet. While being the primary approach in the past, manual collection has now been applied in a limited number of cases for small numbers of large

documents whenever the development of a focused crawler was deemed inefficient. Most of the previously developed corpora within the kernel of the BulNC were compiled manually, such as the Bulgarian "Brown" corpus. Recently, manual compilation has also been used for collecting parallel fiction texts in multiple languages, accounting for 3.70% of the kernel corpus.

3. **Automatic compilation** by web crawling is in general preferred. Some well-known and widely used approaches for automatic collection of corpora are adopted, tailored further to our specific needs and optimised with respect to the efficiency and precision of the output. Currently, automatically obtained subcorpora within the BulNC include a large amount of administrative texts, news from monolingual and multilingual sources, scientific texts and popular science (e.g., Wikipedia articles), altogether amounting to 30.47% of the Bulgarian kernel of BulNC.
   Manual and automatic web mining prior to the crawling process ensures crawling efficiency, as well as high-quality results when it comes to the validity of collected documents and the correspondence between parallel texts. As parallel resources involving Bulgarian are limited on the web, crawling was supported by direct targeting, automatic or manual, of the appropriate resources. The structure of source webpages is also considered when crawling, by applying either links traversal algorithms or URL templates as appropriate for each source.

Several crawling algorithms were examined (Paramita *et al.*, 2011) and the main technique chosen to be applied in the general crawler was the Breadth-First algorithm (Pinkerton, 1994). First, a generalised crawler with the main functionalities was developed. The crawler starts at the initial webpage of the respective collection of documents and either harvests the links recursively until the relevant pages containing the documents are reached, or uses URL templates to access the pages directly. In most cases, the websites containing parallel texts are very large[34] and a general (non-focused) crawler needs to process a very large amount of links and documents in order to select the relevant ones. The general crawler is therefore transformed into a focused crawler by adapting it to the structure of the source site

[34] `http://eur-lex.europa.eu`

as derived by automatic or manual web mining. The focused crawler either implements the link harvesting technique directly, or uses a particular set of URL templates specific for a given website. Next, the focused crawler ensures the relevance of the extracted documents by selecting only those texts that have Bulgarian equivalents. Some corpora are static and require a single run of the crawler, while others are dynamic (e.g., news websites) and need weekly or monthly crawls.

Procedures to verify the validity of the documents collected through automatic crawling are implemented: deletion of empty files obtained from either invalid or missing URLs, text size checks, and verification of encoding. Furthermore, genuine correspondence of parallel documents is checked by comparing URLs, file sizes, dates, etc. To conclude, focused crawling with preceding web structure mining (which considerably reduces the number of visited links) ensures high quality of the results and improves efficiency.

4.2 *Size of the Bulgarian National Corpus*

The kernel of the BulNC, consisting of all Bulgarian texts in the corpus, currently amounts to 979.6 million tokens. Although efforts have been made at ensuring the relative balance of the texts in terms of media type, written texts prevail significantly (91.11%), with spoken data representing only 8.89% of the tokens and being limited in variety – parliamentary proceedings, lectures, and subtitles.

At present, the Bul-X-Cor features 33 parallel corpora, the so-called satellites, adding up to 972.3 million tokens. The kernel and the satellites total 1.95 billion tokens altogether. Each parallel subcorpus within the Bul-X-Cor mirrors the structure of the kernel. Languages are not equally represented: the largest corpus is the Bulgarian-English parallel corpus (280.8 and 283.1 million words for Bulgarian and for English respectively); four other corpora comprise between 100 and 200 million tokens per language, sixteen parallel corpora are in the range of 30 to 52 million tokens per language, another seven in the range of 1 to 10 million tokens, and the rest are below one million, with the smallest ones being the Chinese, the Japanese and the Icelandic corpora with less than 50,000 tokens per language (Figure 2).

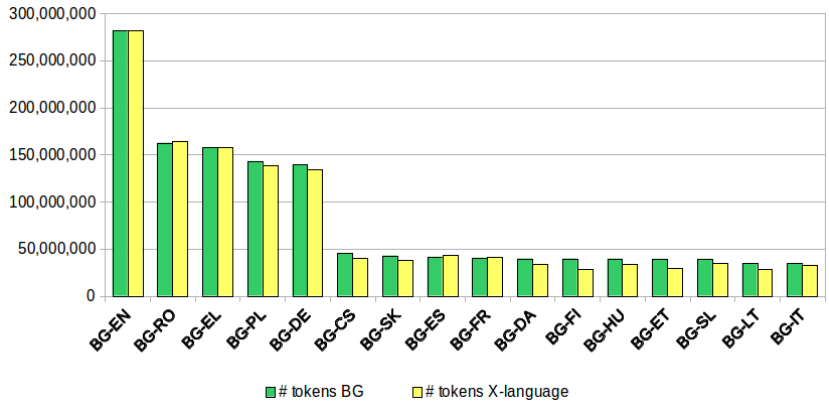


Figure 2: Largest Bulgarian-X language parallel corpora within the BulNC

4.3 *Structure of the Bulgarian National Corpus*

The structure of the corpus adheres to three main principles: explicit definition of categories, clear-cut structure and structure flexibility. The structure is not rigid in the sense that it is not predefined. The corpus samples are supplied with extensive metadata, facilitating the extraction of subcorpora with specific structure and features.

Language reflects communication in the following aspects: function and roles of the participants (style), thematic content (domain), and compositional structure (genre). The realisation of their interconnectivity is essential in building a good model for text description and classification. The design of the corpus is therefore based on the three basic classificatory features of style, domain, and genre.

Style is defined as a general complex text category, which combines the notions of register, mode, and discourse. The proposed approach does not rely on a particular linguistic theory of style, but is based on the analyses of Todorov (1984) and Halliday (1985), among others, who consider the intrinsic characteristics of texts in relation to external, sociolinguistic factors, such as the function of the communicative act.

Different terms are used in the existing literature: speech genre (Todorov, 1984), text type (Biber, 1989), register (Crystal, 1991). We have adopted the term *style* in the sense of Crystal (1969) with a more complex meaning that combines the notion of register (various degrees of formality of language (Trudgill, 1992)), media type (spoken or written) and discourse (function and characteristics of the com-

municative situation as reflected in the text). At present, the BulNC includes texts from six styles. Their distribution measured in number of tokens is presented in Figure 3.

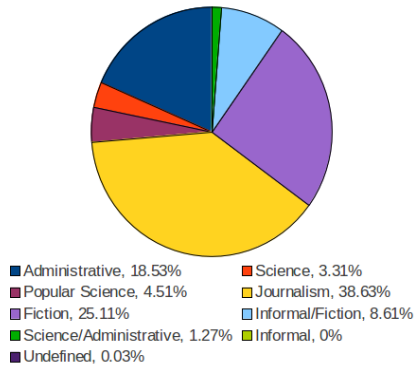


Figure 3:
Distribution of styles in the BulNC

A concise description of the text styles in the BulNC is presented in Table 1. Along with clear-cut styles, two complex styles are also included: Informal/Fiction and Science/Administrative. The former can be defined as informal texts within fiction (subtitles), and the latter as highly specialised (scientific, e.g., medicine) texts within the administrative style. Each of the complex styles exhibits features typical of both components and may share domains and genres with either of them.

Each style is subdivided into thematic domains. It is generally true that domains are style-dependent, although sometimes they are found across styles. For example, the scientific style is divided into categories according to scientific field, e.g., mathematics, economics, political science, etc. Some of the domains of journalistic texts are similar to those of scientific texts – politics, economy, etc.

The term genre also has multiple interpretations. For our purposes we accept the interpretation where genre is associated with the internal formal features of the text (Kress, 1993), although the notion is extended to all texts, both written and spoken. The classification of genres is also inconsistent across linguistic studies, and in particular in existing corpus descriptions (Lee, 2001). A general classification of genres based on style and a set of widely accepted genre types is used in the BulNC.

Table 1: Characteristics of styles in the BulNC

Style	**Communicative situation**	**Function of the text**	**Features of the text**
Administrative	Between official bodies and individual or legal subjects; official, formal, indirect, written	Establishing, regulating and maintaining formal relationships	Relatively strict form and structure, repetitive, ambiguity is avoided
Science	Between researchers and other specialists; formal, indirect, written	Communicating scientific facts	Strict form and structure, extensive use of specialised (domain-specific) language
Popular Science	Between researchers and the wider public; not strictly formal	Communicating scientific facts in accessible and understandable form	Freer form and structure, less specialised language
Journalism	Mainly between journalists and the general public; indirect	Providing information, news and commentary	Relatively stable form and structure, some emphatic elements (e.g., in structure or lexis)
Fiction	Between authors and the general public; indirect	Entertainment and conveying aesthetic and moral values	Free and varied structure, consistent genre-specific elements
Informal	Personal communication; more often direct, informal	Conveying personal message, sharing information	Free and varied structure, diversity in linguistic expression
Informal/ Fiction (Subtitles)	Informal situations within fictional work	Same as fiction; within the fictional framework – personal communication	Characteristics of both styles
Science/ Administrative	Administrative situations within highly specialised scientific domains	Same as administrative	Characteristics of both styles

Style	Number of domains	Number of genres
Administrative	11	16
Science	21	15
Popular Science	25	7
Journalism	19	12
Fiction	13	25
Informal	(not represented)	(not represented)
Informal/Fiction (Subtitles)	17	1
Science/Administrative	21	16

Table 2: Distribution of domains and genres across styles in the BulNC

Table 2 presents the number of domains and genres each style is divided into. Table 3 provides an example of the domains and genres for the Administrative style.

The distribution across domains of the samples in Bul-X-Cor is similar to the distribution in the kernel of the BulNC. The styles are represented as follows:

1. Administrative – EU legislation documents in 23 languages
2. Science/Administrative (Healthcare) – administrative documents from the European Medicines Agency in 23 languages
3. Journalism – news in 9 Balkan languages and English
4. Fiction – texts in Bulgarian, English, German, Romanian, Polish, Greek, Czech.
5. Informal/Fiction – subtitles of feature films, documentaries and cartoons in 29 languages.
6. Science – in Bulgarian and English.

Figure 4 illustrates the distribution of styles within the Bulgarian-English parallel corpus.

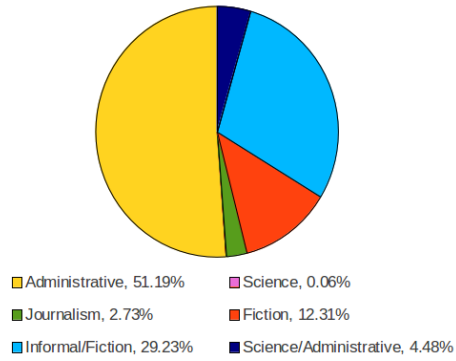
4.4 *Documentation and annotation*

The quality of corpus documentation and annotation has a major impact on the extent of its applications and usability. Therefore, great effort has been made to ensure that the documentation and annotation are accurate, well-structured and compliant with established stan-

Table 3:
Domains and genres for the Administrative style in the BulNC

Domains	**Genres**
Politics	debates
Law	contract
Education	report
Economy	application form
Health	interview
Military	commentary
Culture and arts	correspondence
Sports	law
Ecology	plan
Social policy	programme
Relations	minutes
Undefined	certificate
	directive
	proceedings
	information
	document
	undefined

Figure 4:
Distribution of styles in the Bulgarian-English parallel corpus



dards. Several levels of metadescription have been performed on the texts in the BulNC:

- Metadata documentation – metadata extraction and validation or metadata description of texts in any language;
- Monolingual annotation – processing of texts in any language at various linguistic levels;
- Multilingual annotation – alignment of parallel texts.

4.4.1 Text metadata

The metadata description of the texts in the BulNC is stored into 25 categories (Table 4) that are compliant with the established standards (Atkins, 1992; Burnard, 2005), although defined for the particular needs of the BulNC.

filename	path_to_file	date_added_to_corpus
author_info	author	translator_info
translator	text_info	title
year_of_creation	publishing_date	source_type
source	translated	medium
number_of_words	style	genre
genre_info	domain1	domain2
domain_info	notes	keywords
languages		

Table 4: Metadata in the BulNC

Metadata are mostly derived automatically, using two main techniques: (i) extracting information from the HTML or XML markup of the original files collected from the Internet, and (ii) keyword-based heuristics. HTML pages usually contain specifically tagged editorial information such as author, title, and date of publishing that are easily extractable from the HTML source.

Webpages within a website often contain similar texts, so focused crawling makes it easier to add classificatory information such as domain and genre. When classificatory information is not directly available, heuristics are applied to determine the domain and genre. One very simple example of using lists of domain-specific or genre-descriptive keywords is that if the title of a text contains a genre-specific word (e.g., *report*), it is assumed to denote the genre of the document.

The metadata are as detailed as possible in order to ensure easy text classification, corpus restructuring and evaluation, derivation of subcorpora based on a set of criteria (e.g., year of publication, domain). Some of the metadata categories, labelled with `_info`, are optional and contain additional details about the main category. A multiple domain description was also included to cater for the description of texts which have mixed domain features. So far, extensive metadata are provided for the Bulgarian and the English part of the BulNC, while the corresponding texts from the other languages share the common metadata (author, title, etc.) and inherit the classificatory information for style, domain, and genre.

4.4.2 Monolingual annotation

Until recently, parts of the BulNC, such as the Bulgarian POS-Tagged Corpus, the Bulgarian Sense-Annotated Corpus, the Bulgarian-English Sentence- and Clause-Aligned Corpus were manually annotated (see Section 2.3), while lately we have focused on automatic annotation of larger portions of the corpus.

The linguistic annotation in the BulNC is divided into: (i) general monolingual annotation (tokenisation and sentence splitting), available for all languages, and (ii) detailed monolingual annotation, available only for the languages for which the respective tools and resources are available: Bulgarian and English. The detailed annotation so far includes morphosyntactic tagging (POS tagging and rich morphological annotation), and lemmatisation. The annotation of Bulgarian texts is further extended by including word senses, synonyms, hypernyms and `similar_to` adjectives, noun phrases, and named entities.

The Bulgarian texts are annotated using the Bulgarian language processing chain[35]. It integrates a number of tools (a regular expression-based sentence splitter and tokeniser, an SVM POS-tagger, a dictionary-based lemmatiser, a finite-state chunker, and a wordnet sense-annotation tool), designed to work together and to ensure interoperability, fast performance and high accuracy. The training of the Bulgarian tagger is based on the following parameters: two passes in both directions; a window of five tokens, the currently tagged word being

[35] `http://dcl.bas.bg/dclservices/`

in second position; 2- and 3-grams of words or morphosyntactic tags or ambiguity classes; lexical parameters such as prefixes, suffixes, sentence borders, and capital letters. Lemmatisation is based on linking the tagger output to the Grammatical dictionary (75 word classes to 1029 unique grammatical tags in the dictionary)[36], while a number of rules and preferences are applied to resolve the ambiguities. The finite-state chunker is a rule-based parser working with a manually-crafted grammar designed to recognise unambiguous phrases and to exclude pronouns, adverbs, and relative clauses as modifiers. The context-dependent rules provide annotation for phrase boundaries and heads.

Apache OpenNLP[37] with pre-trained models and Stanford CoreNLP[38] are used for the annotation of the English texts – sentence segmentation, tokenisation, and POS tagging. OpenNLP could be trained and applied for other languages as well. There are also some pre-trained models for a number of widely used languages (German and Spanish, among others). Lemmatisation of the English texts is performed using Stanford CoreNLP and RASP (Briscoe, 2006). As we aim at high quality and consistency of the annotation, we examine various systems for processing English and other languages.

Uniformity in annotation for Bulgarian and other languages is achieved in either of two ways: (i) annotation of raw data from scratch, applying equal standards and principles, or (ii) conversion of already existing annotation. In each case the tagset and conventions accepted for the BulNC are followed. The different tagsets are mapped to the Bulgarian tagset, but any language-specific annotation is preserved. The design of the Bulgarian tagset provides a uniform description of the inflexion of Bulgarian words and multiword expressions (Koeva, 2006) based on morphological and morphosyntactic criteria[39]. The tagset is mappable to the Multext-East morphosyntactic descriptions (Erjavec, 2004; Chiarcos and Erjavec, 2011), which are valuable as a unified framework for many European languages, although some disadvantages have been discovered with regard to the set of descriptions, both on a general and a language-specific level (Przepiórkowski and Woliński, 2003).

[36] `http://dcl.bas.bg/est/dict.php`
[37] `http://incubator.apache.org/opennlp/`
[38] `http://nlp.stanford.edu/software/corenlp.shtml`
[39] `http://dcl.bas.bg/en/BulgarianTagset_en.html`

4.4.3 Multilingual annotation

Multilingual annotation includes alignment at different linguistic levels, currently sentence and clause level. Alignment at sentence level is essential for all parallel resources and it is therefore required for all language pairs. High-quality sentence segmentation is an important prerequisite for the quality of parallel text alignment. The vast majority of the errors that occur in sentence alignment follow from inaccurate sentence segmentation. Two aligners have been applied for parts of the corpus: HunAlign (Varga *et al.*, 2005) and Maligna[40]. The alignment is based on the Gale-Church algorithm, which uses sentence-length distance measure and is largely language-independent. Other alignment methods, such as the Bilingual Sentence Aligner (Moore, 2002) and the use of bilingual dictionaries, are envisaged as well. The aligners take as input texts with segmented sentences and produce a sequence of parallel sentence pairs (bi-sentences). At present, alignment is performed and tested on the Bulgarian-English Parallel Corpus. A further step in parallel corpora processing is automatic clause alignment (Koeva *et al.*, 2012a), currently under way.

4.4.4 Annotation formats

Each raw text in the corpus is in plain text format. The annotation tools exchange data in the so-called vertical format, which is converted into an XML format and then stored in a MySQL database. In the vertical format, the tokens are separated by a newline and the annotation tags by a tab character. Each tool accumulates tags in fixed positions in one or several columns (for tags with a complex structure). Tags can be associated with a single token or a group of tokens.

```
new            TOK_LAT     new          A
technologies   TOK_LAT     technology   Np
.              TOK_FS<S>   .            U
```

The XML format also provides flexibility for representing various levels of annotation in both flat form (as sequences of elements) or hierarchical form (as nested elements, particularly useful for syntactic annotation). In the flat XML format adopted so far the text is represented as a sequence of `words` with associated attributes and their values that store the annotation information.

[40] `http://align.sourceforge.net/`

```
<word w="new" l="new" sen="11439" pos="A">
<word w="technologies" l="technology" sen="11439" pos="Np">
```

4.5 *General evaluation of the BulNC*

The monolingual and parallel parts of the BulNC can be evaluated from several perspectives, either quantitatively or qualitatively.

- **Quality of compilation methods**
  The quality of the crawling is ensured by implementing several techniques: manual and automatic data mining prior to crawling, development of a focused crawler for efficiency, as well as methods for verification of the results.
- **Statistical methods for qualitative analysis and evaluation**
  The strategy to gather a greater variety of word-grams and their distribution rather than to achieve balanced text category distribution is dominant. The aim is to employ statistical methods for qualitative analysis and evaluation, e.g., the proportion between the number of unique tokens / lemmas in the corpus and their frequencies / coverage / distribution within different (combinations of) styles, domains, and genres. It is assumed that variety in word distribution presupposes variety in text categories. For example, sparsity is evaluated through Zipf's law for frequency distribution and type-to-token ratios between old and new words (Goweder and De Roeck, 2001), and violation of Zipf's law may indicate data sparsity.
  Word sequence distribution (higher-order N-grams) may be combined with smoothing and skipping techniques (i.e., calculation of conditional probability based on different context) and with word similarity measures for automatic word clustering (Koeva *et al.*, 2012a). In that respect, the new data not only contribute by adding new lexical units, but also by supporting the saturation of the language model based on the previously collected lexical units.
- **Quality of metadata and annotation**
  Effort is made to ensure high-quality annotation in terms of accuracy, variety and coverage. On the average, in each metadata record in the BulNC 17.79 categories are non-empty (71.16%), a

figure that shows a good overall coverage for the description of the corpus texts.

The POS and grammatical tagger included in the Bulgarian language processing chain (Koeva and Genov, 2011) performs with a precision of 96.58%. The precision reported for the pre-trained model of OpenNLP used for the POS tagging of the English texts is 96.59%. Access to the processing tools is provided through a web service[41] or a web interface for asynchronous tasks[42].

4.6 *Access and applications of the BulNC*

The BulNC has been developed and expanded primarily to meet the needs of natural language processing. Still, the broad range of areas of application of the corpus makes it well-suited for public availability.

4.6.1 Public access to the BulNC

As for the public access to the BulNC[43], we fully comply with Bulgarian and EU legislation concerning copyright and related rights. The law permits the use of copyrighted material for purposes of non-commercial scientific research and for education or private study. Where possible, we extract and store information about the source and the author's name and cite it accordingly. Several types of access to the corpus are provided: (i) download (limited); (ii) web search interface; (iii) collocation service; (iv) subcorpora selection; (v) frequency lists derived from the whole corpus or a given subcorpus.

Due to the inclusion of copyrighted material, the BulNC is not downloadable in full. For several style-specific subcorpora no redistribution limitations are in force, and these are available for download (registration required).

Like many of the large corpora presented in Section 2.1, the BulNC is supplied with a web interface for searching the corpus, as well as for building concordances and extracting collocations. The search system[44] (Figure 5) allows complex linguistic queries involving different levels of annotation combined in various ways. It is designed to support monolingual and parallel corpora in a uniform way. As compared

[41] http://dcl.bas.bg/dclservices/

[42] http://dcl.bas.bg/dclservices/admin/

[43] http://ibl.bas.bg/en/BGNC_access_en.htm

[44] http://search.dcl.bas.bg

to the CQL (Christ and Schulze, 1994), the implemented Designed query language (DQL) (Tinchev *et al.*, 2007) supports terms, such as: word – e.g. `word`; arbitrary word – e.g. `*{POS=A POS=ADV}`, relation – e.g. `word/F/`, and their combinations – e.g. `word/S/{POS=N}`. It is not restricted to a predetermined set of relations – at the moment queries for word forms, synonyms, hypernyms, and `similar_to` adjectives are supported. The atomic formulae allow both ordered and unordered queries, the latter being relevant for matching adjacent constituents with free word order, e.g., verbal clitics in Slavic languages. DQL is recursive and all Boolean combinations of formulae are formulae. This allows, among other things, disjunction of ordered queries, i.e., searching for paraphrases. The system also supports queries with regular expressions. For a given query the system retrieves matches in all documents regardless of language. Thanks to the alignment, the corresponding sentences in parallel documents are also accessible. The hits are paginated and the matches are highlighted. The user is able to view the detailed information for a given sentence in the hit set – the sentence metadata, its context, and correspondence(s) in the other languages.

The BulNC Collocation service employs the free NoSketchEngine[45], a system for corpora processing. The collocation service is a RESTful web service, supporting complex queries through HTTP and providing statistical information.

For instance, the query

```
http://dcl.bas.bg/collocations/?cmd=collocations&word=cat&cbgrfns=3td
```

returns statistical significance calculated with MI3, T-score, and logDice.

In accordance with our view that the corpus should allow for easy compilation of domain- and purpose-specific corpora compliant to a set of predefined criteria – e.g., synchronic, specialised, balanced subcorpora, we intend to provide a web interface for subcorpora selection, processing, and analysis. The extensive metadata ensures that a large set of criteria is available to cater for various research purposes and requirements. At present, we offer an offline request-based service for subcorpora selection and compilation of frequency lists[46].

[45] `http://nlp.fi.muni.cz/trac/noske`

[46] `http://ibl.bas.bg/en/BGNC_access_en.htm`

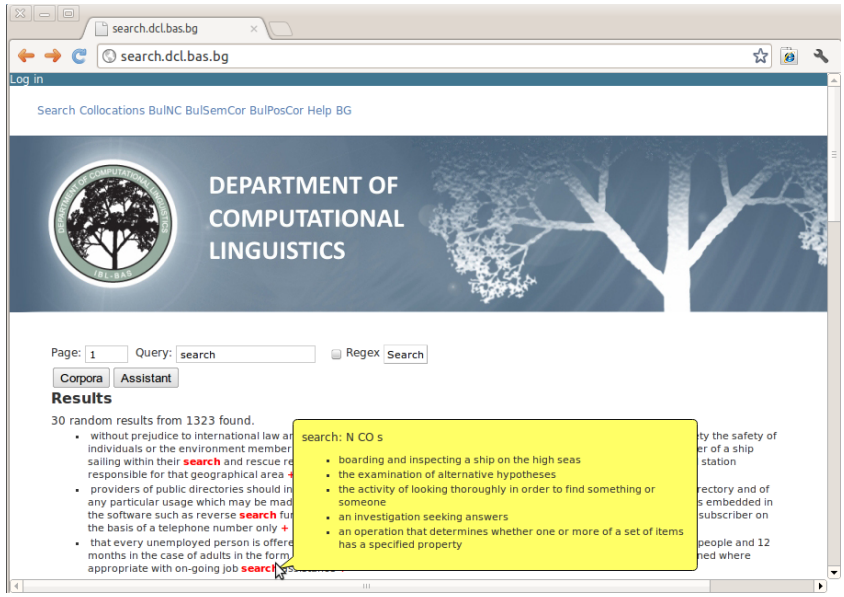


Figure 5: The BulNC search web service

4.6.2 Specialised subcorpora

Manually annotated subcorpora of the BulNC have been used as training and testing resources in numerous studies and NLP tasks, among them theoretical linguistic research, lexicological and lexicographic studies, POS tagging, semantic annotation and disambiguation, MWE recognition, parallel text alignment, clause segmentation and alignment, and many others.

For example, parts of the BulPosCor were used as training and test corpora in the creation of the SVM POS-tagger. The principal application of the BulSemCor is in the training and evaluation of a multi-component word sense disambiguation system. The corpus Wiki1000+, which contains Wikipedia articles (part of the *Popular science* style), includes 13.4 million words. Wiki1000+ was used for the purposes of recognition and classification of multiword expressions. The Bulgarian Sentence- and Clause-Aligned Corpus has been used for the purposes of parallel text alignment at sentence and clause level. It has served as a training resource in the development of a tool for clause alignment (Koeva *et al.*, 2012a). Several Moses[47] models (Koehn and Hoang, 2007) have been built on a large amount of par-

[47] `http://www.statmt.org/moses/`

allel data aligned at the sentence level in order to demonstrate the effect of syntactically enhanced parallel data (clause segmentation and alignment, reordering of clauses, etc.).

The applications of the BulNC and its subcorpora listed here are only a few examples of the numerous applications of the BulNC in the field of natural language processing.

5 CONCLUSION

In the context of the advance of technologies and the fast-growing amount of online information, the notions of text selection, balance, and representativeness can and should be reconsidered, shifting the focus from the theoretically grounded expectation for the distribution of text samples across different domains and genres to more sophisticated and flexible prediction based on calculations and estimations of language usage.

The paper has outlined the main concepts in corpus compilation with an emphasis on the key issues related to the use of corpora for the purposes of NLP research and applications. The attempt at redefining these concepts draws upon a discussion of the principles adopted in the compilation of large monolingual and parallel corpora for various languages. At the present time, large monolingual and multilingual corpora are constructed mostly by amassing text archives, repositories of documents, and bulks of texts available on the Internet.

Against this background, we propose a clear-cut approach for the compilation of a large multilingual corpus and demonstrate it in the context of the Bulgarian National Corpus. Our approach emphasises the extensive metadata and multi-level annotation of very large automatically collected monolingual and multilingual corpora, as well as the uniform treatment of multilingual content with respect to compilation, documentation, annotation, processing, and access.

REFERENCES

Krasimira ALEKSOVA (2000), *Ezikat i semeystvoto. Kam metodikata za prouchvane na rechta v mikroobshtnostite (Language and the family. Towards a methodology for analysis of speech in micro social environments)*, Intervyu Press, Sofia.

Beryl Sue ATKINS (1992), Theoretical lexicography and its relation to dictionary-making, *Dictionaries: Journal of the Dictionary Society of North America*, 14:4–43.

Beryl Sue ATKINS, Jeremy CLEAR, and Nicholas OSTLER (1991), Corpus design criteria, `http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf`.

Michele BANKO and Eric BRILL (2001), Scaling to very very large corpora for natural language disambiguation, in *Proceedings of ACL 2001*, pp. 26–33.

Piotr BAŃSKI, Peter M. FISCHER, Elena FRICK, Erik KETZAN, Marc KUPIETZ, Carsten SCHNOBER, Oliver SCHONEFELD, and Andreas WITT (2012), The New IDS Corpus Analysis Platform: challenges and prospects, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2905–2911.

Piotr BAŃSKI and Adam PRZEPIÓRKOWSKI (2010), The TEI and the NCP: the model and its application, in *Proceedings of LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management (LRSLM2010)*, pp. 34–38.

Marko BARONI and Adam KILGARRIFF (2006), Large linguistically-processed web corpora for multiple languages, in *Proceedings of European ACL, Trento, Italy*, pp. 87–90.

Božo BEKAVAC, Petya OSENOVA, Kiril SIMOV, and Marko TADIĆ (2004), Making monolingual corpora comparable: a case study of Bulgarian and Croatian, in M. T. LINO, M. F. XAVIER, F. FERREIRA, R. COSTA, R. SILVA, C. PEREIRA, F. CARVALHO, M. LOPES, M. CATARINO, and S. BARROS, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, volume IV, pp. 1187–1190.

Douglas BIBER (1989), A typology of English texts, *Linguistics*, 27:3–43.

Douglas BIBER (1993), Representativeness in corpus design, *Literary and Linguistic Computing*, 8(4):243–258.

Ondřej BOJAR, Zdeněk ŽABOKRTSKÝ, Ondřej DUŠEK, Petra GALUŠČÁKOVÁ, Martin MAJLIŠ, David MAREČEK, Jiří MARŠÍK, Michal NOVÁK, Martin POPEL, and Aleš TAMCHYNA (2012), The joy of parallelism with CzEng 1.0, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

Ted BRISCOE (2006), An introduction to tag sequence grammars and the RASP system parser, Technical report, University of Cambridge, Computer Laboratory Technical Report.

Lou BURNARD (2005), *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Metadata for corpus work, Oxford: Oxbow Books, `http://ota.ahds.ac.uk/documents/creating/dlc/index.htm`.

František Čermak and Vera Schmiedtová (2003), The Czech National Corpus Project and lexicography, in M. Murata, S. Yamada, and Y. Tono, editors, *Asialex '03 Tokyo Proceedings: Dictionaries and Language Learning: How Can Dictionaries Help Human and Machine Learning?*, pp. 74–80.

Atanas Chanev, Kiril Simov, Petya Osenova, and Svetoslav Marinov (2006), Dependency conversion and parsing of the BulTreeBank, in *Proceedings of the LREC Workshop Merging and Layering Linguistic Information, Genoa, Italy, 2006*, pp. 16–23.

Jonathan Chevelu, Nelly Barbot, Oliver Boeffard, and Arnaud Delhay (2007), Lagrangian relaxation for optimal corpus design, in *Proceedings of the 6th ISCA Tutorial and Research Workshop on Speech Synthesis (SSW6)*, pp. 211–216.

Christian Chiarcos and Tomaž Erjavec (2011), OWL/DL formalization of the MULTEXT-East morphosyntactic specifications, in *Proceedings of the Linguistic Annotation Workshop 2011*, pp. 11–20.

Oliver Christ and Bruno M. Schulze (1994), *The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual*, University of Stuttgart, Germany.

Jeremy Clear (1992), Corpus sampling, in G. Leitner, editor, *New directions in English language corpora*, Berlin: Mouton de Gruyter.

David Crystal (1969), *What is Linguistics?*, London: Edward Arnold.

David Crystal (1991), *A Dictionary of Linguistics and Phonetics*, Cambridge, MA: Basil Blackwell.

James R. Curran and Miles Osborne (2002), A very very large corpus doesn't always yield reliable estimates, in *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)*, pp. 126–131.

Mark Davies (2010), The corpus of contemporary American English as the First Reliable Monitor Corpus of English, *Literary and Linguistic Computing*, 25(4):447–465.

Ludmila Dimitrova, Tomaž Erjavec, Nancy Ide, Heiki-Jan Kaalep, Vladimir Petkevic, and Dan Tufiş (1998), Multext-East: parallel and comparable corpora and lexicons for six Central and Eastern European languages, in C. Boitet and P. Whitelock, editors, *Proceedings of COLING-ACL'98: 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montréal, Québec, Canada*, pp. 315–319, San Francisco, Calif.: Morgan Kaufmann.

Ludmila Dimitrova, Violetta Koseska, Danuta Roszko, and Roman Roszko (2009), Bulgarian-Polish-Lithuanian Corpus – current development, in *Proceedings of the International Workshop Multilingual resources, technologies and evaluation for Central and Eastern European languages in conjunction with International Conference RANPL 2009, Borovec, Bulgaria, 17 September 2009*, pp. 1–8.

Tomaž Erjavec (2004), MULTEXT-East Version 3: multilingual morphosyntactic specifications, lexicons and corpora, in M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, and S. Barros, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, pp. 1535–1538.

Winthrop Nelson Francis and Henry Kučera (1964), Brown Corpus Manual, `http://icame.uib.no/brown/bcm.html`.

Roger Garside, Geoffrey Leech, and Tony McEnery (1997), *Corpus Annotation: Linguistic Information from Computer Text Corpora*, London: Longman.

Voula Giouli, Nikos Glaros, Kiril Simov, and Petya Osenova (2009), A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts, in *Proceedings of Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, pp. 35–41.

Abduelbaset Goweder and Anne De Roeck (2001), Assessment of a significant Arabic corpus, in *Proceedings Workshop on Arabic Language Processing, 39th ACL, Toulouse*.

Atle Grønn and Irena Marijanovic (2010), Russian in contrast: Form, meaning and parallel corpora, *Oslo Studies in Language (OSLa)*, 2(1):1–24.

Michael Halliday (1985), *An Introduction to Functional Grammar*, Melbourne: Edward Arnold.

Frank Keller and Mirella Lapata (2003), Using the web to obtain frequencies for unseen bigrams, *Computational Linguistics*, 29(3):459–484.

Adam Kilgarriff and Gregory Grefenstette (2003), Introduction to the special Issue on Web as Corpus, *Computational Linguistics*, 29(3):333–347.

Adam Kilgarriff, Vojtěch Kovář, and Pavel Rychlý (2009), Tickbox Lexicography, in *eLexicography in the 21st century: new challenges, new applications*, pp. 411–418, Brussels : Presses universitaires de Louvain.

Jan Kocek, Marie Kopřivová, and Věra Schmiedtová (2000), The Czech National Corpus, in *Proceedings of EURALEX 2000*, pp. 127–132.

Philipp Koehn (2005), Europarl: A parallel corpus for statistical machine translation, in *Proceedings of MT Summit*, pp. 79–86.

Philipp Koehn and Hieu Hoang (2007), Factored Translation Models, in *Proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Prague, Czech Republic, June 2007*.

Svetla Koeva (2006), Inflection morphology of Bulgarian multiword expressions, in *Computer Applications in Slavic Studies*, pp. 201–216, Boyan Penev Publishing Center.

Svetla KOEVA (2010), Balgarskiyat semantichno anotiran korpus – teoretichni postanovki (Bulgarian semantically annotated corpus – theoretical concepts), in *Balgarskiyat semantichno anotiran korpus (Bulgarian semantically annotated corpus)*, IBL.

Svetla KOEVA, Diana BLAGOEVA, and Sia KOLKOVSKA (2010), Bulgarian National Corpus Project, in N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODJIK, S. PIPERIDIS, M. ROSNER, and D. TAPIAS, editors, *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pp. 3678–3684.

Svetla KOEVA and Angel GENOV (2011), Bulgarian language processing chain, in *Proceedings of Integration of multilingual resources and tools in Web applications. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg*.

Svetla KOEVA, Svetlozara LESEVA, Borislav Rizov RIZOV, Ekaterina TARPOMANOVA, Tsvetana DIMITROVA, Hristina KUKOVA, and Maria TODOROVA (2011), Design and development of the Bulgarian sense-annotated corpus, in *Las tecnologías de la información y las comunicaciones: Presente y futuro en el análisis de córpora. Actas del III Congreso Internacional de Lingüística de Corpus. Valencia: Universitat Politécnica de Valéncia*, pp. 143–150.

Svetla KOEVA, Svetlozara LESEVA, Ivelina STOYANOVA, Ekaterina TARPOMANOVA, and Maria TODOROVA (2006), Bulgarian tagged corpora, in *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, Sofia, Bulgaria*, pp. 78–86.

Svetla KOEVA, Borislav RIZOV, Ekaterina TARPOMANOVA, Tsvetana DIMITROVA, Rositsa DEKOVA, Ivelina STOYANOVA, Svetlozara LESEVA, Hristina KUKOVA, and Angel GENOV (2012a), Application of clause alignment for statistical machine translation, in *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6), 12 July 2012, Jeju, Korea*.

Svetla KOEVA, Ivelina STOYANOVA, Rositsa DEKOVA, Borislav RIZOV, and Angel GENOV (2012b), Bulgarian X-language parallel corpus, in N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODJIK, and S. PIPERIDIS, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2480–2486, `http://www.lrec-conf.org/proceedings/lrec2012/pdf/587_Paper.pdf`.

Gunther KRESS (1993), Against arbitrariness, *Discourse and Society*, 4(2):169–191.

Marc KUPIETZ, Cyril BELICA, Holger KEIBEL, and Andreas WITT (2010), The German Reference Corpus DEREKO: A Primordial sample for linguistic research, in N. CALZOLARI, K. CHOUKRI, B. MAEGAARD, J. MARIANI, J. ODJIK, S. PIPERIDIS, M. ROSNER, and D. TAPIAS, editors, *Proceedings of the*

Seventh conference on International Language Resources and Evaluation (LREC 2010), pp. 1848–1854.

David LEE (2001), Genres, registers, text types, domains and style: clarifying the concepts and navigating a path through BNC jungle, *Language Learning & Technology*, 5(3):37–72.

Geoffrey LEECH (1991), The state of the art in corpus linguistics, in *English Corpus Linguistics: Linguistic Studies in Honour of Jan Svartvik*, pp. 8–29, London: Longman.

Christopher MANNING and Hinrich SCHUTZE (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Tony MCENERY, Richard XIAO, and Yukio TONO (2006), *Corpus-Based Language Studies. An Advanced Resource Book*, Routledge.

Charles F. MEYER (2002), *English Corpus Linguistics. An Introduction*, Cambridge University Press.

Robert C. MOORE (2002), Fast and accurate sentence alignment of bilingual corpora, in *AMTA'02: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, pp. 135–144, London, UK: Springer-Verlag.

Graham NEUBIG (2011), The Kyoto Free Translation Task, `http://www.phontron.com/kftt`.

Cvetanka NIKOLOVA (1987), *Chestoten rechnik na balgarskata razgovorna rech (A Frequency Dictionary of Colloquial Bulgarian)*, Sofia: Nauka i izkustvo.

Monica PARAMITA, Ahmet AKER, Robert GAIZAUSKAS, Paul CLOUGH, Emma BARKER, Nikos MASTROPAVLOS, and Dan TUFIŞ (2011), Report on methods for collection of comparable corpora, ACCURAT – Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation.

Brian PINKERTON (1994), Finding what people want: Experiences with the WebCrawler, in *Proceedings of the First World Wide Web Conference, Geneva, Switzerland*, `http://thinkpink.com/bp/webcrawler/www94.html`.

Jan POMIKÁLEK, Miloš JAKUBÍČEK, and Pavel RYCHLÝ (2012), Building a 70 billion word corpus of English from ClueWeb, in N. CALZOLARI, K. CHOUKRI, T. DECLERCK, M. U. DOGAN, B. MAEGAARD, J. MARIANI, J. ODJIK, and S. PIPERIDIS, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 502–506.

Jan POMIKÁLEK, Pavel RYCHLY, and Adam KILGARRIFF (2009), Scaling to billion-plus word corpora, *Advances in Computational Linguistics. Special Issue of Research in Computing Science*, 41:3–14.

Adam PRZEPIÓRKOWSKI (2011), Linguistic annotation of the National Corpus of Polish, FDSL 9, `http://www.uni-goettingen.de/de/document/download/cbcf2e9ded91b3c41d0c460c31d1d9bb.pdf/nkjp.pdf`.

Adam Przepiórkowski, Marek Łaziński, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (2010), Recent developments in the National Corpus of Polish, in N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odjik, and S. Piperidis, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp. 994–997.

Adam Przepiórkowski and Marcin Woliński (2003), The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish, in *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC 2003), EACL 2003*, pp. 109–116.

Kiril Simov and Petya Osenova (2004), A hybrid strategy for regular grammar parsing, in M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, C. Pereira, F. Carvalho, M. Lopes, M. Catarino, and S. Barros, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal*, pp. 431–434.

Kiril Simov, Petya Osenova, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krassimira Ivanova, Alexander Simov, and Milen Kouylekov (2002), Building a linguistically interpreted corpus of Bulgarian: the BulTreeBank, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Canary Islands, Spain*, pp. 1729–1736.

John Sinclair (2005), *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Corpus and text – basic principles, Oxford: Oxbow Books, `http://ahds.ac.uk/linguistic-corpora/`.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Daniel Varga (2006), The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 2142–2147.

Marko Tadić (2002), Building the Croatian National Corpus, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Canary Islands, Spain*, pp. 441–446.

Jörg Tiedemann (2009), News from OPUS – A collection of multilingual parallel corpora with tools and interfaces, in N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pp. 237–248, Amsterdam/Philadelphia: John Benjamins.

Tinko Tinchev, Svetla Koeva, Borislav Rizov, and Nikola Obreshkov (2007), System for advanced search in corpora, in *Literature and Writing in Internet*, pp. 92–111, Sofia: St. Kliment Ohridski University Press.

Tzvetan Todorov (1984), *Mikhail Bakhtin: The Dialogical Principle*, Minneapolis: University of Minnesota Press.

Peter TRUDGILL (1992), *Introducing Language and Society*, London: Penguin.

Dan TUFIŞ, Svetla KOEVA, Tomaž ERJAVEC, Maria GAVRILIDOU, and Cvetana KRSTEV (2009), ID10503 Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages, in *Scientific results of the SEE-ERA.NET Pilot Joint Call, Vienna*, pp. 37–48.

Dániel VARGA, László NÉMETH, P'eter HALÁCSY, András KORNAI, Viktor TRÓN, and Viktor NAGY (2005), Parallel corpora for medium density languages, in *Proceedings of the RANLP 2005*, pp. 590–596.

Ruprecht VON WALDENFELS (2006), Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment, in B. BREHMER, V. ZHDANOVA, and R. ZIMNY, editors, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9*, pp. 123–138, München: Sagner.

Ruprecht VON WALDENFELS (2011), Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB, in Daniela M. and R. GARABÍK, editors, *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20-21 October 2011*, pp. 156–162.

Richard XIAO (2010), Corpus creation, in *The Handbook of Natural Language Processing*, pp. 147–165.

Jia XU and Weiwei SUN (2011), Generating virtual parallel corpus: a compatibility centric method, in *Proceedings of the Machine Translation Summit XIII*.

Dan-Hee YANG, Pascual Cantos GOMEZ, and Mansuk SONG (2000), An Algorithm for Predicting the Relation between Lemmas and Corpus Size, *ETRI Journal*, 22(2):20–31.

Derivational and Semantic Relations of Croatian Verbs

Krešimir Šojat[1], *Matea Srebačić*[2], *and Marko Tadić*[1]
[1] University of Zagreb, Faculty of Humanities and Social Sciences, Zagreb
[2] University of Zagreb, Zagreb

ABSTRACT

This paper deals with certain morphosemantic relations between Croatian verbs and discusses their inclusion in Croatian WordNet. The morphosemantic relations in question are the semantic relations between unprefixed infinitives and their prefixed derivatives. We introduce the criteria for the division of aspectual pairs and further discuss verb prefixation which results in combinations of prefixes and base forms that can vary in terms of meaning from compositional to completely idiosyncratic. The focus is on the regularities in semantic modifications of base forms modified by one prefix. The aim of this procedure is to establish a set of morphosemantic relations based on regular or reoccuring meaning alternations.

Keywords:
derivational morphology, morphosemantic relations, derivational relations, prefixation, semantic relations, Croatian WordNet

1 INTRODUCTION

In this paper we deal with certain types of derivational and semantic relations between Croatian verbs and discusss the possibilities of their inclusion in Croatian WordNet (CroWN), a lexical-semantic net built through the so-called expand model (cf. Vossen, 1998), i.e. by translating and adapting synsets from Princeton WordNet (PWN) into Croatian. At the present time CroWN consists of 10,000 synsets. Approximately 8500 of these are among the basic concept sets of EuroWordNet (EWN) and BalkaNet (BN). Most of the wordnets developed for other languages within these multilingual projects are based on the same or a very similar structure. Each synset in CroWN

was manually translated, adapted and provided with a definition of its meaning and several contextual usage examples. During this time-consuming work it became obvious that differences between Croatian and English are more significant than was initially assumed, especially when dealing with verbs. Despite the fact that PWN is used as a language-independent conceptual structure and as a sort of *interlingua*, some concepts are either not lexicalized in Croatian or are expressed in different parts of speech and therefore do not fit into lexical hierarchies as structured in PWN. For example, the English verb *to face*, as in *The two sofas face each other*, cannot be translated with a single verb in Croatian, and a construction like *to be opposite* should be used instead. Although multi-word literals are used in CroWN, this example and similar ones mainly relating to so-called stative verbs are problematic, since synsets contain words of the same part of speech. Multi-word units are units consisting of two or more words, but the whole unit should be the same POS as the other words in the same synset. However, the construction *to be opposite* consists of a verb and an adjective, and therefore, it is not a multi-word unit belonging to the lexical category of verbs. There are also numerous cases when verbs from PWN have both causative and reflexive translation equivalents in Croatian for the same synset (e.g., the verb *to melt* defined in PWN in one of its senses as *to become or cause to become soft or liquid*). Since all verbs in Croatian are always marked for aspect, the majority of English verbs have two or more translation equivalents in Croatian. The English verbs *drink* and *imbibe* — defined in PWN in one of their senses as *to take in liquids* and contextually illustrated with the sentence *The patient must drink several liters each day* — can be translated with at least five Croatian verbs: *piti, popiti, ispiti, ispijati, poispijati*[1]. Each of these verbs is morphologically derived from the basic verb *piti* 'to drink' through affixation, and each affix differently affects the base form in terms of lexical semantics. Whereas *popiti* denotes only that the action is finished, *ispiti* denotes that the action is finished and there is no liquid left. The lexical meaning of the verb *ispijati* includes both of these components as well as the additional com-

[1] Common strategies in dealing with aspect in Slavic wordnets are described in Section 2.

ponent denoting that the action is performed iteratively. Finally, the verb *poispijati* denotes that all these components are present, but the action is performed either by several subjects or on several objects.

The questions we pose here are (1) how to account for the derivational relations that exist between verbs in Croatian as an inflected language with rich derivation and (2) how to describe these relations in comparison to those which already exists among verbs in CroWN.

2 RELATED WORK

There are six main semantic relations between verbal synsets in PWN: synonymy, antonymy, proper inclusion, troponymy, presupposition and cause (cf. Fellbaum, 1998). As in EWN and BN, troponymy is substituted for hyponymy in CroWN, and cause is extended to encompass presupposition (cf. Vossen, 1998). Whereas the semantic relations in PWN connect synsets consisting of words from the same part of speech, in more recent work Fellbaum et al. (2007) discuss cross-POS relations that hold among words belonging to different synsets sharing a stem with the same meaning. These 14 "morphosemantic links", introduced into PWN 2.0., encompass relations based on suffixation patterns between verbs and nouns. Each relation is semantically labeled (e.g., Agentive, Instrument, Vehicle, Location etc.). None of the morphosemantic links include verb-verb pairs.

Pala and Hlaváčková (2007), Koeva (2008), and Koeva et al. (2008) discuss the problems they faced in the building of Czech, Bulgarian, and Serbian wordnets, respectively. Pala and Hlaváčková (2007) discuss the enrichment of Czech WordNet through the automatic generation of "derivational nests", i.e., new word forms derived from stems by adding affixes associated with specific meanings. They list 14 main derivational processes in Czech between nouns, verbs, adjectives, and adverbs, such as agentive relation in verb-noun pairs or diminutive relation in noun-noun pairs. Relations between derived and base form are semantically labeled and included in Czech WordNet, resulting in a "two-level network". The higher level includes semantic relations between synsets such as synonymy, antonymy, or hyponymy. The lower level includes derivational relations between literals, i.e., single synset members. Verb-verb pairs are linked through

prefixation, but this relation is not taken into consideration in further processing and analysis.[2] Derivational relations are further classified into those which are predominantly semantic in nature, such as *agent*, and those which are predominantly morphological, such as *gerund*.

Koeva (2008) and Koeva et al. (2008) distinguish between morphosemantic and derivational relations. They claim that semantically related synsets in a source language for which the connection has been established through derivation can be used to link the synsets in a target language where such derivational links do not exist in the building of wordnets for several languages inter-connected via e.g., Interlingual index. The former, morphosemantic relations, are not language-specific, whereas the latter, derivational mechanisms of lexicalization, are language-specific. The sharing of semantic information across wordnets on similar grounds has also been proposed by Bilgin (2004) in the building of Turkish WordNet. Koeva (2008) stresses that one of the most productive derivational relations in Bulgarian is between verbal aspectual pairs and points out that perfective and imperfective verbs in Bulgarian WordNet are to be split into separate synsets that are subordinate to the same immediate hypernym. The relation of hypernymy would be based on imperfective verbs only. Derivationally related aspectual verb pairs would therefore be linked as literals. On the other hand, synsets would be linked with the morphosemantic relation *aspect*. The work presented in Koeva et al. (2008) concerning derivational and morphosemantic relations in Serbian WordNet is based on the same grounds and refers mainly to derivational relations across different parts of speech. In Serbian WordNet, aspectual pairs are members of the same synset. Each literal is provided with additional information about its inflectional and aspectual properties. The authors point out that, apart from aspectual pairs, perfective verbs derived from imperfective verbs by prefixation often have a different meaning and thus are not in the same synset.

Extensive accounts of cross-POS derivatives and intra-POS verbal derivatives are given in Maziarz et al. (2011) and Maziarz et al. (2012) for Polish WordNet 2.0. This approach significantly differs from those

[2] Pala and Hlaváčková (2007) stress that "due to a variety of relations that result from combinations of prefixes and base verbs (e.g., distributive, location, time, measure and others), this topic calls for a separate examination (project)".

mentioned above for other Slavic wordnets. Polish WordNet was not developed through the expand model, and its structure heavily relies on lexical units, i.e., literals (word-sense pairs). In Polish WordNet, aspectual pairs are kept apart and lexical hierarchies consist of either perfective or imperfective verbs. Relations between verbs are divided into purely semantic relations (inter-register synonymy, hyponymy and hypernymy, meronymy and holonymy, two types of antonymy, converseness, state, processuality, causality, inchoativity, presupposition, preceding, fuzzynymy) and derivationally-motivated relations (pure aspectuality, secondary aspectuality, iterativity, derivationality, cross-categorial synonymy, role inclusion). Some of the relations hold between lexical units (word-sense pairs, e.g., antonymy or pure aspectuality), while others hold between synsets (e.g., hyponymy and processuality). Although the relations in Polish WordNet do account for verbal derivatives, we think that a more fine-grained analysis is required for Slavic wordnets.

3 DERIVATIONAL RELATIONS BETWEEN CROATIAN VERBS

As in other Slavic languages, Croatian verbs are always marked for aspect and classified as perfective, imperfective, or bi-aspectual.[3] Imperfective and perfective verbs in Croatian can, in terms of derivation, be roughly divided into four groups. The first group comprises non-prefixed imperfective verbs (e.g., *pisati* 'to write'). The second group consists of predominantly perfective verbs built by prefixation of verbs from the first group (e.g., *pre*+*pisati* 'to copy by writing'). The third group comprises imperfective verbs that denote the iterativity of the action. Verbs in this group are built by suffixation of verbs from the second group (e.g., *pre*+*pis*+*iva*+*ti* 'to copy over and over again'). The fourth group consists of perfective verbs derived by prefixation of verbs from the third group (e.g., *is*+*pre*+*pis*+*iva*+*ti* 'to finish copying over and over again'). Verbs in this group include distributive verbs denoting actions performed by several agents usually on several ob-

[3] It is important to stress that so-called bi-aspectual verbs are not imperfective and perfective at the same time. They acquire a perfective or imperfective reading in a particular context.

jects.[4] Aspectual pairs are formed by prefixation or suffixation. Prefixation of an imperfective verb can either yield an imperfective verb (*osjećati* 'to feel' – *su* + *osjećati* 'to sympathize') or a perfective verb (*pisati* 'to write' – *pre* + *pisati* 'to copy'). Prefixation of a perfective verb can yield only perfective forms (*dati* 'to give' – *pre* + *dati* 'to hand over'). Imperfective forms of perfective verbs are built by suffixation (*dati* 'to give' – *da* + *va* + *ti* 'to give repeatedly'). Whereas both types of affixes can be applied to create pure aspectual pairs, only suffixation is used to generate iterative forms of perfective verbs. Aspectual pairs can thus be divided into primary or true aspectual pairs and secondary aspectual pairs. True aspectual pairs are determined primarily by the test of secondary imperfectivization (cf. Raguž, 1997; Jelaska et al., 2005; Maziarz et al., 2011), but also by additional criteria pertaining to the semantics of prefixes. The relation of pure aspectuality exists between a base form and a derivative with a prefix which does not contain any other semantic components except perfectiveness. For example, although the verb *potrčati* 'to start running' does not pass the test of secondary imperfectivization (it is not possible to derive an iterative imperfective **potrčavati*), it is not the pure aspectual pair of the verb *trčati* 'to run$_{\text{ipf}}$' since it additionally denotes the beginning of the action. In other words, the lexical meaning of the verb *potrčati* contains the aspectual component of perfectiveness and the semantic component of inchoativity. The lexical meaning of the verb *dotrčati* 'to run to$_{\text{pf}}$', which denotes the other end of the same action, contains the aspectual component of perfectiveness as well as semantic components of completeness and direction of movement. Therefore, the true or primary aspectual pair of the verb *trčati* is the derived form *otrčati* 'to run$_{\text{pf}}$'[5]. This derived form consists of the prefix *od-* and the imperfective infinitive *trčati* 'to run$_{\text{ipf}}$'. It is a polysemous unit with two different meanings. The first meaning is locative, namely, 'to run from', while the second is 'to finish running'. Due to its second meaning, the derivative *otrčati* is the primary aspectual pair of

[4] Distributive verbs have several subjects (*poiskakati* 'to jump one by one'), several objects (*poubijati* 'to kill one by one') or both several subjects and objects (*pogledavati* 'to glance at each other').

[5] Change from *od* + *trčati* via *ot* + *trčati* to *otrčati* is morphonologically determined (voiced 'd' is devoiced in front of the voiceless 't' and then blended into a single 't').

the verb *trčati*, since the only difference in their meanings is that of aspect. The verb *otrčati* does not pass the test of secondary imperfectivization, and its prefix is regarded as semantically most deprived of its content in comparison to other derivatives. As a general rule we postulate that if a derivational prefix does not have additional semantic components (e.g., inchoativity) apart from the aspectual component of perfectiveness and the derivative does not pass the test of secondary imperfectivization, then the base form and its derivative are true or primary aspectual pairs. Verbs that do not pass the test of secondary imperfectivization, but whose prefixes simultaneously have additional semantic components are not considered true aspectual pairs. This distinction is important since true aspectual pairs are members of the same synset in CroWN. Although the relation between pure aspectual pairs, as well as between secondary aspectual pairs, is considered to be a derivational phenomenon in Croatian linguistics (cf. Barić et al., 2003; Babić, 2002), we treat them as members of same synsets in CroWN since the only difference in meaning between them is the difference in aspect. The same holds for iterative verbs and prefixed perfectives that serve as their base forms. Although iterative verbs have the additional semantic component of repetitiveness, they are grouped in the same synset as their base forms since their lexical meaning is not affected by this additional component. However, the difference in aspect is reflected in the definitions of synset meanings. Each synset member is tagged with one of the following aspect labels: IPF, PF, BI, or ITER, representing imperfective, perfective, bi-aspectual and iterative forms. This distinction is also reflected in different aspectual forms used in definitions that vary according to the aspect of the literals they relate to.[6]

In addition to aspect change, both prefixes and suffixes can create a shift in the meaning of base forms, but the semantic impact of suffixes is rather limited. Apart from the change in aspect, suffixes are used to form verbs denoting diminutive actions and pejorative attitudes. On the other hand, the semantic impact of prefixes is much wider and less predictable. Combinations of prefixes and base forms

[6] Definitions are structurally and semantically the same. The only difference is that imperfectives are defined with imperfective verbs; perfectives with perfectives; bi-aspectual verbs with combinations of imperfective and perfective verbs; and iteratives with imperfectives + 'repeatedly'.

can vary in terms of meaning from compositional to completely idiosyncratic.

4 PREFIXATION

Prefixation is the most productive derivational process of Croatian verbs (Babić, 2002). Verbs are derived by prefixation only from other verbs. There are 19 productive prefixes in Croatian: *do-, iz-, na-, nad-, o-/ob-, obez-, od-, po-, pod-, pre-, pred-, pri-, pro-, raz-, s-, su-, u-, uz-, za-*.[7] The majority of these prefixes are of prepositional origin and have homographic counterparts in prepositions. The prefixes without prepositional counterparts in contemporary Croatian are *obez-*[8]*, pre-, pro-, raz-* and *su-*. Prefixation of Croatian verbs can trigger:

(a) a change in aspect (e.g., *puniti* 'to fill_{ipf}' – *napuniti* 'to fill_{pf}');

(b) a change in aspect and the addition of a new, more specific semantic component to the base form (e.g., *puniti* 'to fill_{ipf}' – *ispuniti* – 'to fill something $\text{completely}_{\text{pf}}$');

(c) the addition of a new component to the meaning of the base form without a change in its aspect (*osjećati* 'to feel_{ipf}'– *suosjećati* 'to $\text{sympathize}_{\text{ipf}}$').

Since prefixes are developed from prepositions, most of them retained their original prepositional meanings. The semantic structure of prefixes can be described as a radial polysemous structure with one central and several peripheral meanings.[9] Very often derivatives acquire two or more of their semantic components. One of them is always more prominent than the others that are nevertheless present in the overall

[7] Apart from these 19 prefixes, there are several non-productive prefixes, as well as prefixes of foreign origin not taken into consideration here.

[8] This prefix is actually a combination of two prepositions (*o* + *bez*), but *bez-* cannot apear as an individual prefix.

[9] Polysemous units as categories with radial structure consisting of a central and several peripheral meanings (determined by metonymical and metaphorical shifts) are presented by Lakoff (1987); Langacker (1987); Raffaelli (2007), and others. Prepositions as polysemous units are analyzed, for example, by Lakoff (1987), and Lindner (1981). For a Slavic analysis of Russian prefixes (cf. Janda, 1985, 1986). Croatian prepositions and prefixes are analysed in the cognitive framework by Šarić (2003, 2006a,b) and Belaj (2008a,b).

semantic structure of derivatives. The polysemous structure of Croatian prefixes can be illustrated with the prefix *za-*, which developed from the preposition *za*. This preposition has several locative meanings, as e.g., *behind* (*Metla je za ormarom* 'There is a broom behind the closet') or *at* (*Sjedi za stolom* 'He sits at the table'). This preposition can be used for expressing temporal relations, as e.g., *during* (*Za vrijeme mog boravka…* 'During my stay…') and *after* (*Dolazi za mnom* 'He will arrive after me'), but also in adverbial constructions of quantity or manner. The semantic complexity of the preposition *za* is also reflected in the polysemous structure of the prefix *za-*. As other prefixes, *za-* can be used for the derivation of pure aspectual pairs, but also for the production of derivatives with specific meanings, e.g.:

(a) locative meaning
    (1) to put something behind something (*zabaciti* 'to throw behind')
    (2) to put something onto something (*zakačiti* 'to attach')
    (3) to change position (*zaleći* 'to lie down')
    (4) to move around (*zakrenuti* 'to go in a curve or around the corner')
(b) inchoative meaning (*zapjevati* 'to start singing')
(c) more or less intensified action (*zamisliti se* 'to ponder', *zagorjeti* 'to scorch')
(d) change of property (*zacrvenjeti se* 'to become red')
(e) pure aspectuality (*zaklati* 'to slaughter')

Despite such complex semantic structure, we believe that the specific meanings of the prefix *za-*, apart from pure perfective meaning in (e), can be divided into four larger classes which we label: (1) location, (2) quantity, (3) time, and (4) manner. In a similar analysis applied to other productive prefixes mentioned above, we tried to deduct their meaning components and their impact on the meaning of the derived verbs. The resulting combinations of prefixes are divided into two groups: (1) pure aspectual pairs and (2) secondary aspectual pairs (cf. Figure 1). The first one is described above and we shall not go into further details.

The group of secondary aspectual pairs is further divided into two broader classes according to the semantic criterion: (1) compositional

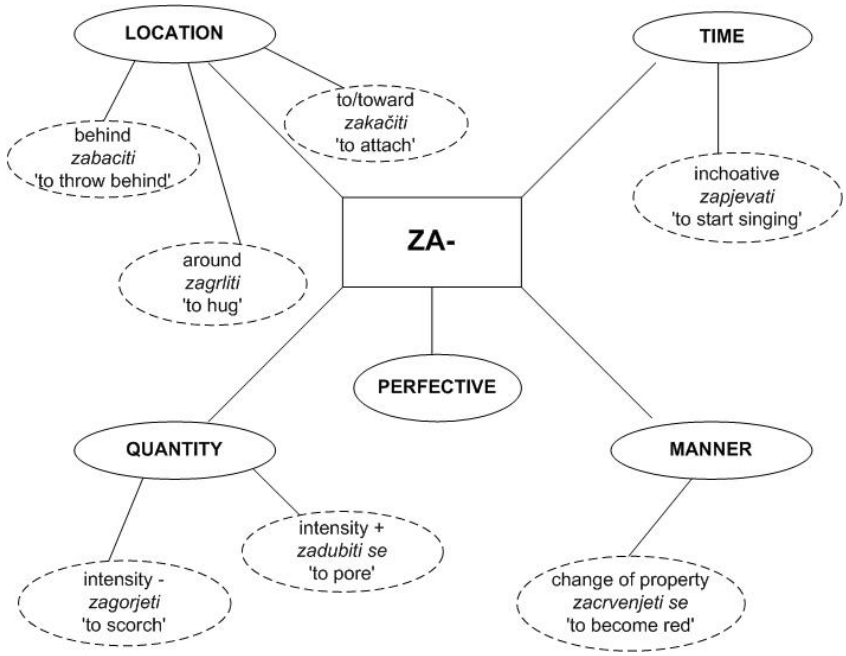


Figure 1: Meanings of the prefix *za-*

and (2) idiosyncratic. The division is motivated by the extent of the semantic shift that takes place in derived forms. Combinations of prefixes and base verbs in Croatian form a continuum in terms of semantic compositionality. On one pole of this continuum there are compositional combinations, i.e., one of the specific meanings of a prefix and lexical meaning of a verb are semantically transparent (e.g., *govoriti* 'to speak' – *progovoriti* 'to start speaking', *pjevati* 'to sing' – *zapjevati* 'to start singing'). On the other pole of this continuum there are completely idiosyncratic combinations. In these combinations the meaning of the derivatives as a whole cannot be directly connected either to the meaning of the prefix or to the lexical meaning of the verb without a thorough analysis of metaphorical or metonymical shifts (e.g., *baciti* 'to throw' – *pobaciti* 'to abort pregnancy'; *pustiti* 'to release' – *napustiti* 'to abandon').

In further sections we focus on predominantly compositional combinations. The goal of this research is to detect and describe meanings of prefixes that are constant and present in combinations with base verbs, i.e., those prefixal meanings that occur even when attached to

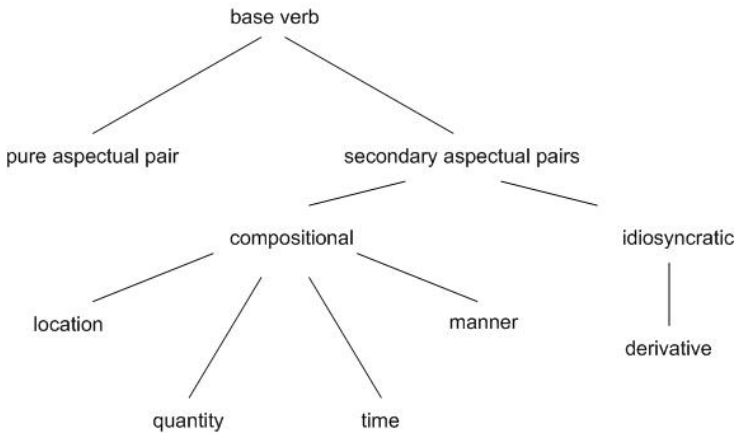


Figure 2:
Derivationally motivated relations between base verbs and derivatives

verbs from various semantic fields.[10] The objective of this procedure is first to determine the set of prefixal meanings that reoccur in various semantic fields and secondly to determine which prefixes can carry the same meanings. The final objective is to establish the set of derivationally motivated semantic relations between Croatian verbs. We will further refer to these relations as *morphosemantic relations*. These are further analyzed in order to determine which relations should be introduced into Croatian WordNet, since they are not encompassed by the existing semantic relations.

To fulfill these tasks, it is necessary to determine which prefixes take part in the derivation of particular base forms. The data on the derivational spans of verbs so far have not been systematically and extensively presented in Croatian morphology. In other words, large-scale data indicating which affixes are used or can be used with particular base forms in Croatian do not exist.

4.1 *Derivational Database*

In order to address these issues, we have collected approximately 14,000 verbal lemmas from digital and freely available dictionaries of Croatian. The initial list consisted of infinitives unsorted in any way. The verbs from the list were automatically processed using a rule-based approach. In the first step of processing we applied a set of rules

[10] Verbs are divided into 15 semantic fields in PWN (cf. Fellbaum, 1998). The semantic fields were taken from WordNet 1.5 (so-called “lexicographic files”) and mapped onto verbal synsets in CroWN.

for the segmentation of prefixes in order to obtain base forms and their derivatives formed through prefixation. The set of rules was designed to remove the 19 productive prefixes presented in the Section 4, as well as their combinations. In Croatian, one base form can bear one, two, three and very rarely even four derivational prefixes at the same time (e.g., $pre_{\text{prefix1}} + ras_{\text{prefix2}} + po_{\text{prefix3}} + dijeliti_{\text{base-form}}$ 'to reassign, to reallocate'). On the other hand, only one derivational suffix is added to roots. Besides a derivational suffix, stems can have either zero or one conjugational suffix before the infinitive endings *-ti* or *-ći*. Conjugational suffixes indicate verbal inflectional classes. The second set of rules was created to recognize and segment the suffixes and the roots. The verb *izrezuckati* 'to cut into small pieces$_{\text{pf}}$' was thus segmented into a prefix (*iz-*), a root (*-rez-*), a derivational suffix (*-uck-*), a conjugational suffix (*-a-*), and an infinitive ending (*-ti*). A form without any derivational affix, i.e. the base form of this derivative, is the verb $rezati_{\text{ipf}}$. The aim of this procedure was also to obtain a set of base forms that are either non-prefixed infinitives, i.e. lemmas, or morphological stems. These morphological stems are not lemmas, although they are used in further derivational processes. For example, the stem **laziti* can acquire different prefixes and thus serves as the base form for the derivation of verbs such as *do-laziti* 'to come$_{\text{ipf}}$', *iz-laziti* 'to exit$_{\text{ipf}}$', *pre-laziti* 'to cross$_{\text{ipf}}$' etc., but it cannot stand alone as an individual word.

In numerous cases, the rules could not detect an affix due to graphical overlapping with its stem. Prefixes were not accurately detached when they were graphically identical to parts of stems. For example, verbs like *privilegirati* 'to privilege' or *sniježiti* 'to snow' were incorrectly segmented as **pri + vileg + ir + a + ti* instead of *priv-ileg + ir + a + ti* and **s + nijež + i + ti* instead of *snijež + i + ti*. A similar problem occurred with suffixes and roots. For example, *krijumčariti* 'to smuggle' was segmented into *krijum + čar + i + ti* instead of *kri-jumčar + ø + i + ti* and pobjeći 'to run away' into *po + b + ø + je + ći* instead of *po + bje + ø + ø + ći*. The output of processing was therefore manually checked and corrected. The final result of this rule-based semiautomatic procedure is a derivational database consisting of infinitives segmented into lexical and grammatical morphemes. This database has enabled further exploration of the derivational network of verbs sharing the same base form. A sample of the database is given in Figure 3. Each lexical entry in the derivational database consists

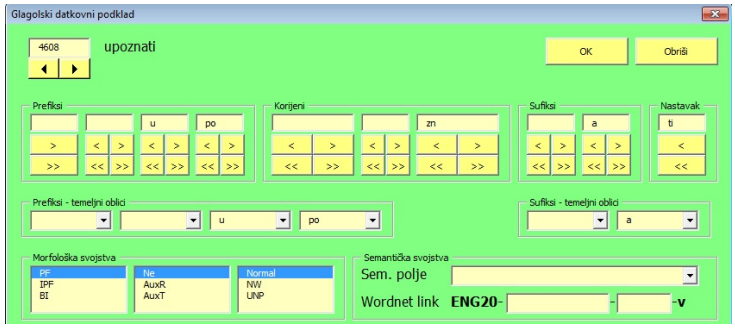


Figure 3:
Sample of derivational database of Croatian verbs

of verbs decomposed into groups of morphemes, and each group of morphemes is provided with slots for roots, affixes and linguistic meta-data. Slots for morphemes are divided into: 1. derivational prefixes (four slots)[11], 2. the lexical part (three slots – in the majority of cases only one slot is filled, the three slots are provided for verbal compounds of two lexical morphemes and an interfix), 3. derivational and conjugational suffixes (two slots) 4. infinitive ending (one slot). The meta-data in lexical entries indicates verbal aspect, types of reflexivity, etc. The database enables queries across the full derivational span of particular base forms and generalizations regarding the distribution and frequency of affixes in the derivation of verbs from other verbs and verbal stems. In its present shape the database comprises 16,834 entries consisting of 14,291 lemmas and 2543 productive stems, i.e. the stems used the in formation of at least 2 derivatives.[12] In the remainder of the paper we focus on combinations of one prefix and a base form, the most productive derivational process among Croatian verbs according to the data from the database. The database has enabled the recognition of 4221 unprefixed base forms used in the prefixal derivation of 10,070 verbs. The distribution of prefixes across slots in the database is given in Figure 4 (P1 – the first slot next to the root contains a prefix, P2 – the first and the second slot next to the root contain prefixes etc.).

[11] Combinations of various prefixes differently affect the meaning of the base form. The combinations and derivations possible from these derived forms are the subject of further research. Preliminary results are shown in Šojat et al. (2012).

[12] The derivational database will be further expanded in terms of other parts of speech. Queries over the database will be possible through a web interface, which is still under construction.

Figure 4: Distribution of prefixes across slots in derivational database

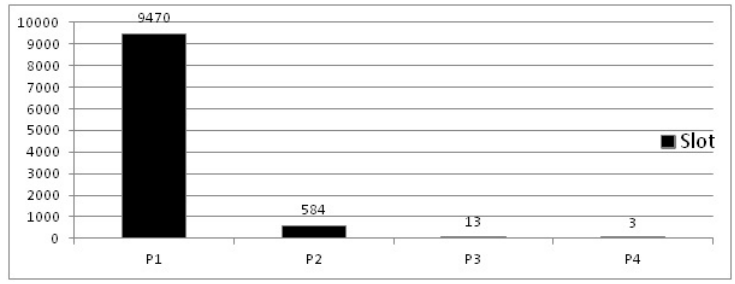
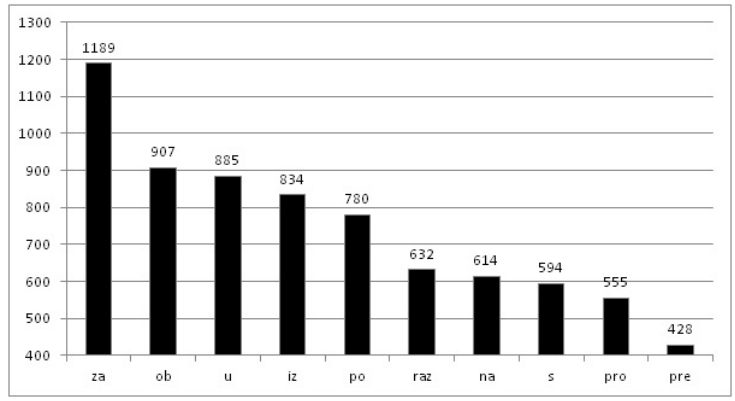


Figure 5: Frequency of particular prefixes in the derivational database: 10 most frequent prefixes in the P1 slot



The ten most frequent prefixes in such combinations are given in Figure 5. The recognition of possible and attested combinations of prefixes and base forms has enabled the classification of prefixal meanings into broad and general categories mentioned in Section 4 above. These categories serve as a basis for further analysis and elaboration of the morphosemantic relations between verbal base forms and verbal derivatives.

4.2 *Prefixal Meanings*

Prefixes in Croatian usually have various and heterogeneous meanings which are often hard to capture and to separate from one another within the same prefix. As mentioned, we have focused on predominantly compositional combinations of prefixes and base forms. In order to establish the set of morphosemantic relations among verbs we have divided prefixal meanings into four major groups: (1) location, (2) time, (3) quantity and (4) manner. These four major groups were further divided into several subgroups. The division into four major

groups is built upon already existing categorizations of prefixal meanings in Croatian grammars (cf. Babić, 2002; Barić et al., 2003) and a preliminary systematization for the purposes of introducing morphosemantic relations into CroWN (cf. Srebačić, 2011). In this paper we have further divided the four major groups into several subgroups upon more in-depth analysis of the prefixal meanings, presented below.

In the *location* group, the prepositional origin of prefixes pervades in their meaning and is more prominent than in the other groups. In this group, prefixes primarily denote spatial relations, i.e., a particular direction or location. The *time* group includes prefixal meanings that refer to various phases of the action denoted by base verbs, such as the beginning or the termination of the denoted action. The *quantity* group includes prefixal meanings that refer to amount or intensity of the action as determined by a prefix. Finally, the *manner* group includes prefixal meanings related to various modes of action denoted by base verbs. Table 1 lists all 19 prefixes and all their meanings established in the analysis and used in further processing.

Such a thorough analysis of 19 productive prefixes in Croatian and their meanings enabled the establishment of major groups and subgroups of morphosemantic relations, which will be presented in the following section.

5 MORPHOSEMANTIC RELATIONS

Labels for morphosemantic relations consist of two parts. The first one pertains to one of the four major groups: location, time, quantity, and manner. The second part indicates particular subgroups of major groups.

location_

1. loc_bott_up – upward movement
2. loc_top_down – downward movement
3. loc_prox – movement in proximity to a subject or object
4. loc_through – movement through something (or someone)
5. loc_apart – movement in opposite or multiple directions
6. loc_to_toward – movement to or toward something or someone

Table 1: The meanings of verbal prefixes

Prefix	Location	Time	Quantity	Manner
do-	1. **to/toward** – *doletjeti* 'to fly to'	1. **completion** – *dozreti* 'to ripen', *dostaviti* 'to deliver' 2. **finitiveness** – *domisliti se* 'to think out'	1. **addition** – dodati 'to add', dopisati 'to add in writing'	
iz-	1. **bottom-up** – *izrasti* 'to grow up', *izroniti* 'to emerge' 2. **from** – *izletjeti* 'to fly from', *izliti* 'to pour out'	1. **distributivity** – *izbacati* 'to throw out one by one' , *ispisati* 'to print' 2. **completion** – *izliječiti* 'to cure'	1. **sufficiency** – **isplakati se** 'to cry one's eyes out', *izvikati se* 'to shout one's fill' 2. **excessiveness** – *izmučiti se* 'to exhaust oneself'	
na-	1. **top-down** – *nabosti* 'to prick, to spike' 2. **proximity** – *naići* 'to come across' 3. **to/toward** – *nalijepiti* 'to paste'	1. **inchoativity** – *natrunuti* 'to begin to rot', *nagristi* 'to start to bite/corode' 2. **distributivity** – *navoziti* 'to cart one by one'	1. **sufficiency** – *najesti se* 'to stuff oneself' 2. **excessiveness** – *napiti se* 'to get drunk' 3. **intensity** – *naraditi se, namučiti se* 'to tire oneself out with work', *nagorjeti* 'to scorch' 4. **addition** – *naloviti* 'catch a quantity of something'	
nad-	1. **over** – *nadgraditi* 'to outbuild'; *nadletjeti* 'to fly over'		1.**exceeding** – *nadrasti* 'to outgrow', *nadjačati* 'to overpower'	
o-/ob-	1. **around** – *okružiti* 'to encircle'; *oploviti* 'to circumnavigate, to sail around'			

obez-			1. **deprivation** – *obezvrijediti* 'to devaluate'	
od-	1. **apart** – *odletjeti* 'to fly away', otići 'to leave'	1. **completion** – *odigrati* 'to play', *odsvirati* 'to play a musical piece'		
po-	1. **top-down** – *poleći* 'to lay down', *posoliti* 'to salt'	1. **inchoativity** – *potrčati* 'to start running', *poletjeti* 'to start flying' 2. **distributivity** – *pomrijeti* 'to die one by one', *pobiti* 'to kill one by one'	1. **intensity** – *poprati* 'to wash a little', *poigrati se* 'to play a little'	
pod-	1. **under** – *podbosti* 'to spur', *podložiti* 'to place under'		1. **insufficiency** – *potplatiti* 'to underpay', *pothraniti* 'to feed insufficiently'	
pre-	1. **over** – *preskočiti* 'to jump over', *preletjeti* 'to fly over' 2. **re-location** – *preseliti* 'to relocate', *pretočiti* 'to pour over'	1. **completion** – *prenoćiti* 'to spend the night'	1. **intensity** – *presoliti* 'to oversalt', *pregrijati* 'to overheat' 2. **exceeding** – *prerasti* 'to outgrow'	1. **change of property** – *pretvoriti se* 'to convert', *preimenovati* 'to rename'
pred-		1. **preceding** – *pretplatiti se* 'to pay in advance', *prethoditi* 'to precede'		
pri-	1. **proximity** – *primaknuti se* 'to come closer', 2. **to/toward** – *prikačiti* 'to attach', *pribiti* 'to pin down'		1. **intensity** – *primiriti se* 'to calm down a little' 2. **addition** – *priliti* 'to add by pouring'	1. **connection** – *prišiti* 'to sew on'
pro-	1. **through** – *probiti* 'to break through', 2. **proximity** – *projuriti* 'to pass quickly by', *prohujati* 'to rush by'	1. **inchoativity** – *progovoriti* 'to start talking' 2. **completion** – *prožvakati* 'to finish chewing' 3. **preceding** – *proreći* 'to predict'	1. **intensity** – *prodrmati* 'to shake a little', *proprati* 'to rinse'	

raz-	1. **apart** – *razdvojiti se* 'to separate', *raširiti se* 'to spread'		1. **intensity** – *razljutiti se* 'to become very angry'	
s-	1. **top-down** – *srušiti* 'to knock down, to fell', *sletjeti* 'to land'			1. **connection** – *spojiti* 'to bond, to bring together'
su-	1. **proximity** – *susresti se* 'to meet', *sudariti se* 'to bump'			1. **connection** – *sufinancirati* 'to cofinance' 2. **opposition** – *sučeliti se* 'to face'
u-	1. **into** – *uplivati* 'to swim into', *urasti* 'to grow into'	1. **finitiveness** – *ugaziti* 'to trample'	1. **intensity** – *usjedjeti se* 'to sit for a long time', *uznojiti se* 'to sweat abundantly'	1. **change of property** – *usmrdjeti se* 'to become stinky', *uprljati se* 'to become dirty'
uz-	1. **proximity** – *uspinjati se* 'to climb', *uzdizati se* 'to ascend'	1. **inchoativity** – *uskomešati se* 'to stir up'	1. **intensity** – *uzburkati* 'to stir up', *ushodati se* 'to walk up and down'	
za-	1. **around** – *zagrliti* 'to hug' 2. **behind** – *zabaciti* 'to throw back' 3. **to/toward** – *zakačiti* 'to attach' 4. **top-down** – *zaleći* 'to lie down'	1. **inchoativity** – *zatrčati se* 'to start running', *zapjevati* 'to start singing'	1. **intensity** – *zadubiti se* 'to pore', *zagorjeti* 'to scorch'	1. **change of property** – *zacrveniti se* 'to become red'

7. loc_over – movement over something or someone
8. loc_into – movement into something (or someone)
9. loc_around – movement around something or someone
10. loc_under – movement or location beneath something or someone
11. loc_reloc – movement to another location
12. loc_behind – movement behind something or someone
13. loc_across – movement across something
14. loc_from – movement away from something or someone

This group predominantly consists of verbs of movement, since various spatial relations are inherent in their lexical meanings. These

Table 2: Morphosemantic relations in *location* group

Location	Prefix
bottom-up – *uspeti se* 'to climb', *izrasti* 'to grow up'	iz-, po-, uz-
top-down – *porušiti* 'to pull down' , *nabosti* 'to spike', *sletjeti* 'to land'	na-, po-, s-, za-
proximity – *naići* 'to come across', *približiti se* 'to come closer', *projuriti*	na-, pri-, pro-, su-
through – *probiti* 'to break through', *prošiti* 'to quilt'	pro-
apart – *odvojiti* 'to separate', *otkinuti* 'to detach'	od-, raz-
to/towards – *prikačiti* 'to attach', *zabiti* 'to nail', *nalijepiti* 'to stick'	na-, pri-, za-
over – *natkriti* 'to cover over', *preskočiti* 'to jump over'	nad-, pre-
into – *utrčati* 'to run into', *urasti* 'to grow into'	u-
around – *okružiti* 'to circle', *obletjeti* 'to fly around something', *obuhvatiti* 'to embrace'	o-/ob-, za-
under – *podrediti* 'to subject', *podložiti* 'to place under'	pod-
re-location – *preliti* 'to decant', *preseliti* 'to move'	pre-
behind – *zabaciti* 'to throw back'	uz-, za-
across – *prijeći* 'to cross', *preletjeti* 'to fly over', *preplivati* 'to swim across'	pre-
from – *izletjeti* 'to fly from', *izliti* 'to pour out'	iz-

relations also hold between numerous base verbs and their derivatives from other semantic fields, e.g., *prošiti* 'to quilt', *preliti* 'to pour over'. Due to their prepositional origin, prefixes primarily denote spatial relations. For this reason, the majority of prefixes have at least one meaning corresponding to one of the location relations. This fact in turn results in a rather extensive set of morphosemantic relations of location. All location morphosemantic relations with examples are listed in Table 2.

time_

1. time_inch – beginning of the action ('to start X'[13])
2. time_fin – termination of the action ('to finish X')

[13] X = base verb.

Table 3: Morphosemantic relations in *time* group

Time	Prefix
inchoativity – *pojuriti* 'to start rushing', *zaplivati* 'to start swimming', *prozboriti* 'to start talking'	na-, po-, pro-, uz-, za-
finitiveness – *doletjeti* 'to fly to', *dotrčati* 'to run to'	do-, na-, u-
distributivity – *izdijeliti* 'to give one by one', *popadati* 'to fall one by one'	iz-, na-, po-
preceding – *pretkazati* 'to predict', *prethoditi* 'to forego', *pretplatiti* 'to subscribe'	na-, pred-, pro-

3. time_distr – the action performed by several subjects usually on several objects and in successive phases ('repeatedly X')
4. time_prec – the action denoted by derivatives precedes the action denoted by base verbs[14]

The group of time relations is determined by aspectual properties and constraints of Croatian verbs. Relations in this group do not hold between pure aspectual pairs. Besides the aspectual difference between imperfective base forms and perfective derivatives, derivatives also denote various phases or temporal components of the denoted action, such as its starting or terminative point. Morphosemantic relations belonging to the time group with examples are in Table 3.

quan_

1. quan_suff – the action denoted by the derivative is performed in sufficient or insufficient quantity ('enough/not enough X')
2. quan_exc – the action denoted by the derivative is performed in excessive quantity ('too much X')
3. quan_int – the action denoted by the derivative is performed with weaker or stronger intensity ('X a little/a lot')
4. quan_more – the action denoted by the derivative outperforms the action denoted by the base verb that is performed by one or more different subjects ('X better than')

[14] The pure semantic relation *preceding* exists in Polish WordNet (cf. Maziarz et al., 2011), where this relation is used between synsets. In our approach, this relation holds between derivationally related verbs.

Quantity	Prefix
sufficiency (+/−) – *istrčati se* 'to run enough'; *potplatiti* 'to underpay', *pothraniti* 'to feed insufficiently'	iz-, na-, pod-
excessiveness – *napiti se* 'to get drunk', *prejesti se* 'to gormandize', *izmučiti se* 'to exhaust oneself'	iz-, na-, pre-
intensitiy (+/−) – *nagristi* 'to bite a little', *protresti* 'to shake a little', *ustrčati se* 'to bustle around', *razbjesniti se* 'to become very furious'	na-, po-, pre-, pri-, raz-, u-, uz-, za-
exceeding – *nadigrati* 'to outplay', *nadrasti* 'to outgrow'	nad-, pre-
deprivation – *obeshrabriti* 'to discourage', *obezbojiti* 'to decolour'	obez-
addition – *dogrijati* 'to heat to the desirable degree', *dopisati* 'to add by writing'	do-, na-

Table 4: Morphosemantic relations in *quantity* group

5. quan_depr – the action denoted by the derivative refers to the loss of property[15]
6. quan_add – the action denoted by the derivative refers to the addition or completion of the action denoted by the base verb ('to add by X-ing')

As far as we know, quantity as a morphosemantic category has not been accounted for in related work or lexical resources. Our analysis has shown, however, that it must be taken into consideration when dealing with the prefixation of Croatian verbs and the morphosemantic relations between base forms and derivatives. Moreover, since it comprises six subgroups, we firmly believe this group is well justified. Morphosemantic quantitative relations with examples are listed in Table 4.

mann_

1. mann_conn – the action denoted by the derivative refers to two or more inter-related entities. This relation comprises *connection* and *opposition* as stated in Table 1.

[15] Although properties are generally expressed by adjectives, there are verbs derived from adjectives denoting the same property. This relation holds between such verbs and their verbal derivatives (e.g., *obeshrabriti* 'to discourage' is derived from the base verb *hrabriti* 'to encourage', which is in turn derived from the adjective *hrabar* 'courageous'.

Table 5: Morphosemantic relations in *manner* group

Manner	Prefix
inter-connection – *sudjelovati* 'to co-participate', *prikrpati* 'to sew on by patching', *sjediniti* 'to compound'	pri-, s-, su-
change of property – *zazelenjeti se* 'to become green', *ukiseliti se* 'to become sour'	o-/ob-, po-, pre-, s-, u-, za-

2. mann_prop – the action denoted by the derivative refers to the acquisition of property denoted by the base verb

The *manner* group consists of only two subgroups, but these specific meanings cannot be subsumed by any other major groups of relations. Each subgroup of relations is expressed by three or more different prefixes, forming a rather coherent and delimited group of meaning components. *Manner* morphosemantic relations with examples are in Table 5.

We also came across numerous derivatives that cannot be directly connected to their base verbs via any of the listed morphosemantic relations. As mentioned, there are combinations of prefixes and base verbs that are completely idiosyncratic. We mark the relation between such verbs with the underspecified relation *derivative* (cf. Maziarz et al., 2011).

5.1 *Morphosemantic relations in CroWN*

As mentioned above, our final objective was to establish the set of morphosemantic relations between Croatian verbs and determine which relations should be introduced into Croatian WordNet since they are not encompassed by the existing semantic relations.

CroWN contains 2318 verbal synsets with an average of 5.8 verbs per synset. Each verbal synset consists of verbs marked for their senses (so-called *literals*). The total number of verb senses, i.e. literals, is 13,476.[16] For example, *dati* ' to give$_{\text{ipf}}$' is marked for 28 senses and thus appears in 28 different synsets and *letjeti* 'to fly$_{\text{ipf}}$' is marked for 9 senses and appears in 9 different synsets.[17] There are 13 derivatives of

[16] PWN 3.0. comprises 25,047 verbal literals divided into 13,767 verbal synsets. Although the number of synsets in CroWN may seem rather small in comparison to the number of verbal synsets in PWN, the number of verb literals in CroWN is ca. 50% of the number of verb literals in PWN.

[17] Such a particularization of meaning is a consequence of the adopted expand

the verb *letjeti* formed with 8 different prefixes. In only one case does it occur that this base form and a derivative are aspectual pairs and therefore members of the same synset. The remaining 8 base forms and 12 derivatives are members of different synsets, and in more than 50% of cases they are members of different lexical hierarchies based on the semantic relation of hyponymy. In other words, *letjeti* 'to fly_{ipf}' is not positioned in the same hierarchy as the verb *uletjeti* 'to $\text{fly into}_{\text{pf}}$', even though they differ only in the meaning component 'moving into something'. In CroWN we use the same semantic relations between verbal synsets as in EWN and BN. These relations are synonymy, hyponymy/hyperonymy, antonymy, cause, and subevent. The relation of hypernymy/hyponymy is the most important for the overall structure of the lexicon. This relation can be described as 'to do X in a particular manner', where X is a hypernym. For example, verbs of movement are divided into several subfields on the basis of their specific meaning properties, such as manner of movement, direction of movement, the medium in which the movement is performed, means of movement, etc. Such a division results in hierarchies containing heterogeneous groups of hyponyms connected to their co-hypernym only through this specific meaning component. For example, 'to move' has hyponyms in the subfield of direction such as 'to move upwards', 'to move downwards', 'to move across something', 'to move through something', 'to move over something', and 'to move around something'. These hyponymy subclasses contain verbs denoting different media of movement, vehicles, manner, speed, etc. Apart from the general similarity that pertains to the concept of movement, verbs in these subclasses have significantly different meanings and frequently share only one meaning component, e.g., 'to move into something' (*uplivati* 'to $\text{swim into}_{\text{pf}}$', *utrčati* 'to $\text{run into}_{\text{pf}}$', *uletjeti* 'to $\text{fly into}_{\text{pf}}$') or 'to start moving' (*potrčat*i 'to $\text{start running}_{\text{pf}}$', *zaplivati* 'to $\text{start swimming}_{\text{pf}}$', *poletjeti* 'to $\text{start flying}_{\text{pf}}$'). This in turn results in hierarchies that do not contain derivationally related verbs that sometimes differ only in this particular meaning component. Therefore, we have proposed a set of relations between derivationally related verbs which are usually scattered across different hierarchies. In order to determine which morphose-

model (cf. Section 1) and in many cases does not truly reflect semantic structure and relations between Croatian verbs.

mantic relations could or should be introduced between base forms and derivatives in CroWN we conducted an experiment consisting of several steps.

In the first step we removed the sense tags from the literals and reduced this list to single appearances of forms. In other words, literals as tokens were treated as morphological types, resulting in 5747 unique forms. This list was automatically filtered for those verbs containing combinations of 2 and 3 prefixes, verbs with derivational suffixes, and iterative verbs formed by conjugational suffixes (cf. Section 3). The filtering was done by matching this list with the data from the derivational database (cf. Section 4). The output was a list of 2530 base forms and derivatives with only one prefix. This list was further filtered for 754 derivatives marked as aspectual pairs in CroWN. Finally, we obtained a list of 1922 verbal types in CroWN. These forms were used in the second step of the experiment. In this step we segmented prefixed forms into prefixes and base forms, again matching them with the derivational database. Thus we obtained 572 base forms and 1350 derivatives as candidates for the assignment of established morphosemantic relations (cf. Section 5). In the final step, we automatically assigned morphosemantic relations, according to particular prefixes as listed in tables 2-5, to each derivative and manually checked the results. In this analysis we either: (1) eliminated all suggested relations when none of them was appropriate due to the idiosyncratic nature of the combinations and tagged them as DERIV (cf. Figure 2) or (2) we chose the appropriate relation from the total of suggested relations. The result of the whole procedure is a list of 572 base forms and 1204 prefixed verbs marked for morphosemantic relations as described above. The distribution of morphosemantic relations according to particular prefixes and their overall frequency is given in Table 6.

The overall statistics concerning the four major groups of morphosemantic relations and their subgroups, as well as the number of occurrences between base forms and derivatives from CroWN is given in Tables 7, 8, 9[18] and 10.

[18] Two morphosemantic relations marked by * in the quantity table do not occur between verbs in CroWN, although they do occur between verbs in the derivational database. This is due to the significantly smaller number of base formes and derivatives in CroWN than in the derivational database.

Table 6:
Prefixes and morphosemantic links

Prefix (overall freq.)	**Morphosemantic relation**	**Freq.**
do- (35)	loc_to_toward	21
	time_fin	11
	quan_add	3
iz- (133)	loc_from	62
	time_fin	42
	quan_exc	10
	time_distr	10
	quan_suff	5
	loc_bott_up	44
na- (71)	loc_top_down	12
	quan_int	12
	quan_add	11
	loc_to_toward	8
	time_inch	7
	quan_exc	5
	time_distr	2
	time_fin	2
	loc_across	1
	loc_prox	1
	time_prec	1
o-/ob- (83)	loc_around	71
	mann_prop	12
od- (88)	loc_apart	75
	time_fin	13
po- (108)	quan_int	41
	loc_bot_up	25
	time_inch	21
	time_distr	16
	loc_top_down	3
	mann_prop	2
pod- (10)	loc_under	5
	quan_suff	5
pre- (61)	loc_over	26
	mann_prop	14
	quan_exc	11
	loc_reloc	5
	quan_int	2
	time_fin	2
	loc_across	1

Prefix (overall freq.)	Morphosemantic relation	Freq.
pred- (4)	time_prec	4
pri- (66)	loc_to_toward	34
	quan_int	12
	loc_prox	10
	mann_conn	5
	quan_add	5
pro- (81)	loc_through	34
	time_fin	20
	quan_int	11
	time_inch	11
	loc_prox	3
	time_prec	2
s- (59)	mann_conn	27
	loc_top_down	26
	mann_prop	6
su- (6)	mann_conn	4
	loc_prox	2
u- (61)	loc_into	61
	time_fin	40
	mann_prop	28
	quan_int	3
uz- (32)	quan_int	8
	loc_prox	6
	time_inch	6
	loc_behind	2
za- (140)	time_inch	68
	mann_prop	27
	quan_int	16
	loc_around	15
	loc_to_toward	7
	loc_top_down	4
	loc_behind	1

Group (overall freq.)	Subgroup	Freq.
LOCATION (600)	loc_apart	**141**
	loc_araond	87
	loc_from	70
	loc_to_toward	70
	loc_top_down	68
	loc_into	60
	loc_through	34
	loc_over	24
	loc_prox	23
	loc_bott_up	6
	loc_under	6
	loc_behind	5
	loc_reloc	4
	loc_across	2

Table 7:
Frequency of morphosemantic links – *location* group

Group (overall freq.)	Subgroup	Freq.
TIME (276)	time_fin	**132**
	time_inch	109
	time_distr	28
	time_prec	7

Table 8:
Frequency of morphosemantic links – *time* group

Group (overall freq.)	Subgroup	Freq.
QUANTITY (190)	quan_int	**126**
	quan_exc	25
	quan_suff	20
	quan_add	19
	quan_depr	0*
	quan_more	0*

Table 9:
Frequency of morphosemantic links – *quantity* group

Group (overall freq.)	Subgroup	Freq.
MANNER (122)	mann_prop	**88**
	mann_conn	34

Table 10:
Frequency of morphosemantic links – *manner* group

6 DISCUSSION

As mentioned in Section 5.1, the semantic relations between verbal synsets in CroWN are hypernymy/hyponymy, synonymy, antonymy, cause, and subevent. All of them hold between whole synsets and none of them holds between particular literals, i.e., base verbs and their derivatives. Sometimes base verbs and derivatives are connected via one of the semantic relations that holds between synsets. In these cases, base verbs and derivatives are members of different synsets. However, in the majority of cases, morphosemantic and semantic relations do not overlap. Moreover, none of our morphosemantic relations can be completely subsumed by any of these semantic relations in terms of their semantic content.

As far as the semantic relation of hypernymy/hyponymy is concerned, we have indicated that base verbs and their derivatives are often not members of same lexical hierarchies and thus close semantic relations resulting from derivational processes are not recognizable between them.

Antonymy exists between two derivatives of the same base (e.g., *doći* 'to arrive' – *otići* 'to leave'), but this relation does not exist between derivatives and a base form (*ići* 'to go' – *doći* 'to arrive' and *ići* 'to go' – *otići* 'to leave'). The semantic relation cause holds between synsets and denotes the relation between two actions, the first denoting the cause and the second denoting the result or the consequence of action denoted by the first verb (e.g., *hraniti* 'to feed' – *jesti* 'to eat' or *ciljati* 'to aim' – *pogoditi* 'to shoot'). Although the relation cause semantically partially overlaps with our morphosemantic relation change of property, cause can only encompass pairs such as *kiseliti* 'to pickle$_{\text{ipf}}$' – *ukiseliti* 'to pickle$_{\text{pf}}$', but not their reflexive counterparts denoting the non-agentive action, e.g., *kiseliti se* 'to become sour$_{\text{ipf}}$' – *ukiseliti se* 'to become sour$_{\text{pf}}$'. The relation of subevent in EWN and BN denotes the relation between two synsets referring to two simultaneous actions or to an action which is a part of the action denoted by another synset (e.g., 'to eat' has subevents 'to chew' and 'to swallow'). This relation does not refer to derivationally related literals and does not reflect particular parts of events, e.g., its beginning or terminating point, as our morphosemantic relations of inchoativity or finitiveness do.

Since we have marked all of the verbs from CroWN that have one prefix (approx. 30% of all verbs in CroWN) with morphosemantic or aspectual relations, and only two of our morphosemantic relations were not applied at all, we believe that our inventory of relations is well justified and applicable not only in CroWN, but also in wordnets for other Slavic languages. From the related work on other Slavic languages presented in Sections 2 and 4 (especially Footnote 7), it is clear that the same interplay between base verbs and derivatives regarding the semantic impact of prefixes can be found in all branches of Slavic languages. We are convinced that this is always the case in South Slavic languages and that it also holds for Czech, Polish, and Russian.

The problem of including derivational relations in Slavic wordnets is already recognized and discussed, as shown in Section 2. However, the solutions presented do not seem to be fine-grained enough to include all morphosemantic relations between verbal derivatives. Based upon our analysis of prefixal meanings and their classification, we believe that the notion of secondary aspectuality can be further analyzed and divided into at least four major subgroups: *time*, *location*, *quantity* and *manner*. We strongly believe that these four major groups of secondary aspectuality, due to the similarity of Slavic languages, can be applied to other Slavic wordnets without significant changes. The morphosemantic subrelations presented here are probably more language-specific, and their existence or possible implementation should be examined for each Slavic language, but we believe that most of them can be applied to other Slavic wordnets.

7 CONCLUSION

Since all the relations in Croatian WordNet hold between synsets and not between single verbs, so far it has not been possible to account for morphosemantic relations between base forms and their derivatives as described above. The same problem has been detected for other Slavic wordnets, but the presented solutions are not fine-grained enough.

The work done on the derivational database of Croatian verbs has enabled the restructuring of the relations between verbs in CroWN, their adaptation to the lexical properties of the Croatian language and

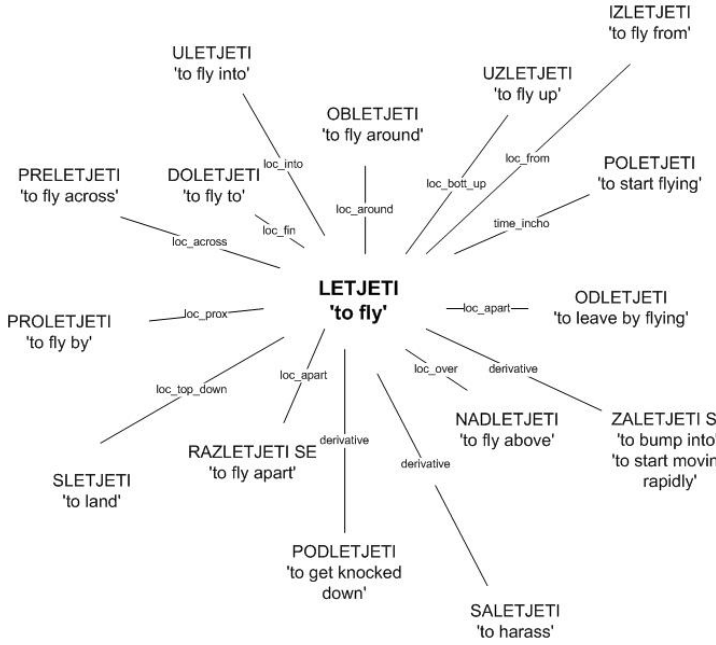


Figure 6: Base form *letjeti* 'to fly' and its derivatives with meanings and morphosemantic relations

the enrichment of CroWN with morphosemantic relations as presented above.

The morphosemantic relations discussed here, resulting from combinations of one prefix and base forms, were first divided into four major groups and further into several subgroups. Combinations of multiple prefixes with the same base form and their influence on lexical meaning have yet to be investigated. This could potentially lead to a further expansion of the morphosemantic relations as stated here.

REFERENCES

Stjepan Babić (2002), *Tvorba riječi u hrvatskome književnom jeziku*, Zagreb: HAZU: Nakladni zavod Globus.

Eugenija Barić et al.(2003), *Hrvatska gramatika*, Zagreb: Školska knjiga.

Branimir Belaj (2008a), *Jezik, prostor i konceptualizacija. Shematična značenja hrvatskih glagolskih prefiksa*, Osijek: Filozofski fakultet.

Branimir Belaj (2008b), Pre-locativity as the schematic meaning of the Croatian verbal prefix pred-, *Jezikoslovlje*, 9(1-2):123–140.

Orhan BILGIN, Özlem ÇETINOGLU and Kemal OFLAZER (2004), Building a WordNet for Turkish, *Romanian Journal of Information Science and Technology*, 7(1-2):163–172.

Bernard COMRIE (1989), *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems*, Cambridge: Cambridge University Press.

Östen DAHL (1985), *Tense and aspect systems*, Oxford: Basil Blackwell.

Christiane FELLBAUM ed. (1998), *WordNet: An Electronic Lexical Database*, Cambridge: MA: MIT Press.

Christiane FELLBAUM, Anne OSHERSON and Peter E. CLARK (2007), Putting Semantics into WordNet's "Morphosemantic" Links, in *Proceedings of the Third Language and Technology Conference*, Poznan (Poland).

Laura JANDA (1985), The meaning of Russian Verbal Prefixes: Semantics and Grammar, Flier, A. S., Timberlake, A., eds. *The scope of Slavic aspect (UCLA Slavic Studies)*, Columbus, Ohio: Slavica, 26–40.

Laura JANDA (1986), *A Semantic Analysis of the Russian Verbal Prefixes ZA-, PERE-, DO- and OT-*, München: Otto Sagner.

Laura JANDA (2004), A metaphor in search of a source domain: the categories of Slavic aspect, *Cognitive Linguistics*, 15(4):471–527.

Zrinka JELASKA (2005), *Hrvatski kao drugi i strani jezik*, Zagreb: Hrvatska sveučilišna naklada.

Svetla KOEVA (2008a), Derivational and Morphosemantic Relations in Bulgarian WordNet, in *Proceedings of the Intelligent Information Systems 2008*, 359–368.

Svetla KOEVA, Cvetana KRSTEV and Duško VITAS (2008), Morpho-semantic Relations in WordNet – a Case Study for two Slavic Languages, in *Proceedings of the 4th Global WordNet Conference*, 239–254.

George LAKOFF (1987), *Women, Fire and Dangerous Things. What Categories Reveal about the Mind*, Chicago&London: The University of Chicago Press.

Ronald W. LANGACKER (1987), *Foundations of Cognitive Grammar. Vol. 1*, Stanford: Stanford University Press.

Susan J. LINDNER (1981), *A Lexico-Semantic Analysis of English Verb-Particle Constructions with UP and OUT*, PhD's dissertation, University of California, San Diego.

Marek MAZIARZ et al.(2011), Semantic Relations between Verbs in Polish WordNet 2.0, *Cognitive studies*, 11:183–200.

Marek MAZIARZ, Maciej PIASECKI and Stan SZPAKOWICZ (2012), An Implementation of a System of Verb Relations in plWordNet 2.0, in *Proceedings of the 6th Global WordNet Conference*, 181–188.

Karel PALA and Dana HLAVÁČKOVÁ (2007), Derivational Relations in Czech WordNet, in *Proceedings of the Workshop on Balto-Slavonic Languages*, 75–81.

Maciej PIASECKI, Stan SZPAKOWICZ and Bartosz BRODA (2009), *A Wordnet from the Ground Up*, Wrocław University of Technology Press.

Ida RAFFAELLI (2007), Neka načela ustroja polisemnih leksema, *Filologija*, 48:135–172.

Ida RAFFAELLI et al.(2008), Building Croatian WordNet, in *Proceedings of the 4th Global WordNet Conference*, 349–360.

Dragutin RAGUŽ (1997), *Praktična hrvatska gramatika*, Zagreb: Medicinska naklada.

Borislav RIZOV (2008), Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 1523–1528.

Ljiljana ŠARIĆ (2003), Prepositional categories and prototypes: Contrasting some Russian, Slovenian, Croatian and Polish examples, *Jezikoslovlje*, 4(2):187–204.

Ljiljana ŠARIĆ (2006a), A preliminary semantic analysis of the Croatian preposition u and its Slavic equivalents, *Jezikoslovlje*, 7(1-2):1–43.

Ljiljana ŠARIĆ (2006b), On the meaning and prototype of the preposition pri and the locative case: A comparative study of Slavic usage with emphasis on Croatian, *Rasprave instituta za hrvatski jezik i jezikoslovlje*, 32:225–248.

Krešimir ŠOJAT, Nives MIKELIĆ-PRERADOVIĆ and Marko TADIĆ (2012), Generation of Verbal Stems in Derivationally Rich Language, in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 928–933.

Matea SREBAČIĆ (2011), *Morphosemantic description of verbs of change in CroWN*, MA thesis, Department of Linguistics, Faculty of Humanities and Social Sciences, University of Zagreb.

Piek VOSSEN ed. (1998), *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*, Dordrecht: Boston: London: Kluwer Academic Publishers.

Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech

György Szaszák[1] *and András Beke*[2]
[1] Department of Telecommunication and Media Informatics,
Budapest University for Technology and Economics, Budapest, Hungary
[2] Research Institute for Linguistics,
Hungarian Academy of Sciences, Budapest, Hungary

ABSTRACT

The relation between syntax and prosody is evident, even if the prosodic structure cannot be directly mapped to the syntactic one and vice versa. Syntax-to-prosody mapping is widely used in text-to-speech applications, but prosody-to-syntax mapping is mostly missing from automatic speech recognition/understanding systems. This paper presents an experiment towards filling this gap and evaluating whether a HMM-based automatic prosodic segmentation tool can be used to support the reconstruction of the syntactic structure directly from speech. Results show that up to 85% of syntactic clause boundaries and up to about 70% of embedded syntactic phrase boundaries could be identified based on the detection of phonological phrases. Recall rates do not depend further on syntactic layering, in other words, whether the phrase is multiply embedded or not. Clause boundaries can be well assigned to intonational phrase level in read speech and can be well separated from lower level syntactic phrases based on the type of the aligned phonological phrase(s). These findings can be exploited in speech understanding systems, allowing for the recovery of the skeleton of the syntactic structure, based purely on the speech signal.

Keywords: *prosody, syntax, phonological phrase, boundary detection*

1 INTRODUCTION

A number of applications in automatic speech understanding require some analysis of the content prior or parallel to speech-to-text conversion referred to often as automatic speech recognition. In speech understanding, a pure transcription of the speech yielded by a speech-to-text converter (speech recognizer) would be insufficient, as the underlying meaning remains unextracted, uninterpreted. Of course, text-based analysers can be used for the speech-to-text output to assess meaning, however, this output can be unreliable depending on the difficulty of the speech recognition task, closely linked to several factors like environmental ones (noises in the speech signal, distortions), or speaking style (the "spontaneity" of speech) or in general the complexity of the recognition task (vocabulary, language model perplexity), etc. For some languages – like Hungarian, which is in the center of interest of the present study – both speech recognition (Szarvas et al., 2000) and text-based syntactical analysis (Babarczy et al., 2005) are difficult and work with significantly weaker performance compared to the English baselines due to the very rich morphology of the language. If recognition performance is poor, only highly unreliable data could be fed into a text-based syntactic or semantic analyser to assess the meaning. On the other side, if speech recognition works with good accuracy for a given task, the interpretation of the meaning can be still supported or constrained by other methods than text-based analysis, in order to add redundancy and create a more robust and more powerful system.

The speech signal itself carries information related to syntax, represented by speech prosody. This means that syntax and prosody interact, even if they cannot be mapped directly and unambiguously to each other (Selkirk, 2001). From a linguistic point of view, the majority of theories dealing with the syntax-prosody relationship conclude that syntax and prosody are closely related, but this relation cannot be expressed as a definite mapping between syntax and prosody. The *prosodic structure hypothesis* by Selkirk (2001) postulates that the prosodic structure of a sentence is related to (but not fully dependent on) the surface syntactic structure. In contrast, some theories argue that prosody is directly governed by the surface syntactic phrase structure (Kaisse, 1985), but evidence shows rather that the relationship syntax-

prosody is more difficult, especially as we are approaching the lower levels (or layers – the two terms are used as synonyms throughout this paper) of the prosodic hierarchy.

Approaching the same problem from point of view of human perception, several studies have proven that prosody is an important clue in syntactic parsing and that prosody constrains lexical access (Cristophe et al., 2004), which proves its essential role in human speech perception. Imaging techniques tracing human brain activity during speech perception by ERP (Event Related Potential) or PET (Positron Emission Tomography) measurements also support this hypothesis (Li-Yang-Lu, 2010), and it is suspected that prosody is a predictive clue for syntactic (and semantic) processing in human perception, justified by ERP tests allowing for the tracing of brain activity (Strelnikov et al., 2006). Indeed, brain areas situated in the dorsolateral prefrontal cortex, associated with the perception of prosody are very close to (or rather overlapping with) those responsible for syntactic analysis, forming a real prosody/syntax interface in the human brain (Strelnikov et al., 2006).

Evidence and practice in speech technology also shows the practical usefulness of the prosody/syntax interface. In text-to-speech conversion, syntactic analysis of a written sentence has become a common task prior to speech synthesis (Koutny-Olaszy-Olaszi, 2000), (Becker et al., 2006). The first initiatives date back even to the 1980s. The underlying assumption for this approach is that the required prosodic features of the sentence to be synthesized can be well predicted relying on syntactic analysis. In other words, one assumes that surface syntactic structure determines the prosodic structure. As this determination is only partial, applications were often restricted to well described domains, such as name and address synthesis (Silverman, 1993).

Despite the fact that the relation between syntax and prosody has widely been exploited in text-to-speech synthesis (Hirschberg, 1993), it is not explicitly included in speech recognition: although prosody is implicitly modelled in *segmental domain* (i.e a domain proportional to the length of a phoneme) through energy related features and implicit duration modelling, it remains the most often neglected in *suprasegmental domain* (i.e. the length of a word or group of words) as explained in Vicsi-Szaszák (2010). Some other studies have already also highlighted this technological gap (Batliner et al., 2006) long time

ago. Since then, several attempts were made to integrate prosody into speech recognition and understanding, focusing essentially on boundary detection tasks using an event detection like approach (Veilleux-Ostendorf , 1993), (Gallwitz al., 2002), (Shriberg et al., 2000). Iwano (1999) and Vicsi-Szaszák (2010) implemented alignment based segmentation and boundary detection upon this segmentation. Going one step further and mapping prosody to syntax and deducing some syntactical attributes based on prosody or perform disambiguation is even less frequently used, although is not completely unknown in speech understanding (Price et al., 1991), (Nöth et al., 2000). A fully statistical approach for information extraction based on prosody was presented by Shriberg-Stolcke (2004), without using labelled corpora to train machine learning based structural and pragmatic taggers, speaker and word recognizers.

These considerations lead and motivate us to experiment with – at least partial – recovery of the syntactic structure in speech based on prosody, an important clue when carrying out automatic interpretation of the content encoded in the speech signal, in order to assess its meaning. Applications so far in this domain are almost exclusively dedicated to disambiguation problems (Price et al., 1991), (Nöth et al., 2000), where prosody is used to select between ambiguous hypotheses (minimal pair sentences) given a speech sample and its different interpretations represented by different syntactic structures. The selection between minimal sentence pairs can be a realistic problem in automatized dialogues, however, it does not provide a globally useful framework to assess syntax based on prosody. Our goal is to fill this gap and implement and test a more globally applicable framework (in contrast to the work presented by Price et al. (1991)) for syntactic analysis based only on speech, capable of providing more detailed analysis based also on training with hand labelled data (in contrast to the work presented by Shriberg-Stolcke (2004)). The outcome of this activity can be useful in several technologies involving speech understanding, like dialogue-based automatized systems with speech interface, automatic interpretation (speech translation), and in general, in any application where analysis of meaning (focus detection, topic detection, keyword spotting, speech segmentation based on prosody, syntactic or semantic analysis, etc.) is crucial.

Instead of creating a more or less artificial corpus of minimal pair sentences, which usually provides only a moderate and often non-realistic corpus for analyzing purposes, a general, large speech corpus is used. The main interest is to analyse the nature of the relation between automatically performed prosodic analysis (phonological phrase boundary detection and classification) and automatically generated and previously disambiguated syntactic analysis (this latter represents the surface syntactic structure). The basic interest is to explore and evaluate to what extent different levels (called also layers (Selkirk, 2001)) in the prosodic and syntactic hierarchy can be mapped to each other, and to analyse further if any type of phonological or syntactic phrase exists which has a special impact on syntactic or phonological structure, respectively. An important side-outcome of the experiment is to evaluate to what extent automatic prosodic segmentation for phonological phrases can reflect the underlying syntactic structure.

Experiments to be presented in this paper were carried out for the Hungarian language, however, special emphasis is also put on the universality and possible extension of the approach for other languages.

This paper is organized as follows: First, syntactic analysis issues are revised, then the prosodic segmentation of speech is presented in details. Hereafter, experiments and results are presented for the reconstruction of the syntax based on speech prosody, followed by conclusions.

2 SYNTACTIC ANALYSIS

The syntactic analysis is performed on the transcripts of utterances which will be used for the evaluation of prosody-to-syntax mapping experiments. The syntactic analysis – provided as a syntactic phrasing – serves as a reference when correspondence of the prosodic and syntactic structure is investigated. First, some basic Hungarian specificities are briefly presented, necessary for the comprehension of the syntactic analysis method used, which has to deal with rich morphology and relatively free word order. Syntactic analysis is presented next, with a short outlook explaining the necessary morphological considerations.

2.1 *Specificities of Hungarian syntax*

Hungarian is an agglutinating language, with a very rich morphology, and consequently, grammatical relations are expressed less by word order constraints and more by suffixes. This allows also for a relatively free word order, where word ordering is more submitted to the fine semantic tuning of the meaning, as case information is available via the suffixes. For example, in English and in many other languages a basic sentence would start with the subject (noun), followed by the verb, ended by the object (noun). In Hungarian, the object is differentiated by the objective case (usually suffix -t) and hence is identifiable as object even if it is moved within the sentence.

Hungarian sentences can be divided into a topic and a predicate (or comment) part (É. Kiss, 2002). The topic part either contains constituents whose denotation (normally, an individual) counts as given in the context, or those denoting entities, properties or eventualities constituting new information that are intended to be contrasted to their alternatives. In sentences with a narrow or contrastive focus, the focused unit is placed between the topic and the verb and must directly precede the latter – this is the focus position. In other words, constituents before the verb are associated with specific functions, whereas units following the verb do not normally express new information (comment part).

Given that Hungarian is an agglutinating language, i.e. grammatical information is expressed by suffixes rather than word order, the primary role of word order is to express information structure.

(a) Mária ismeri Józsefet.

(b) Józsefet Mária ismeri.

Thus, the utterance “Mary (Mária) knows Joseph (József)” can take the forms given in (a) and (b) without a semantic change (Mary is the subject and Joseph the direct object in both sentences), but the information structure is different: sentence (a) has either broad focus with an accent on all content words (also called either a neutral sentence or a sentence with verbal accent), or Mary is in focus and hence the accented unit in the sentence (verbs are deaccented if the focus position is filled). In the latter case, the sentence could be an answer to the question “Who knows Joseph?”. In sentence (b), Mary is in fo-

cus, but the word order indicates that this sentence is about Joseph (= the topic) and includes the option of contrastivity (i.e. "… but it is Rebecca who knows Isaac").

2.2 *Syntactic phrasing*

The syntactic analyser is a language-dependent tool (however, of course, its output is standardized as used in automatic machine translation tools for example). As experiments presented in this paper were done for the Hungarian language, the freely available Hunpars tool (Babarczy et al., 2005) was used as a syntactic analyser was used. This syntactic analyser uses a so-called *phrase-structure grammar*, completed by *lexical databases* and a *morphological analyser* to perform syntactic analysis of written sentences. The analyser outputs tagged and layered syntactic analysis hypotheses for each input sentence.

In a phrase structure grammar, words are grouped into syntactic phrases, which together form a hierarchic (or layered) structure (Gazdar et al., 1985). The identification of grammatical dependencies follows based on this hierarchical grouping. The syntactic phrasal structure is output by bracketing, preserving the hierarchy so that it can be easily converted into a tree-like representation (see an example in Fig. 1).

The phrase structure grammar used in Hunpars is *head-driven* (Pollard-Sag, 1994): each syntactic phrase has a head, corresponding to the word that determines the behavior of the phrase within the syntactic constituent (embedding syntactic phrase or the sentence) located one level up in the hierarchy. For example, the syntactic phrase 'a főkonzul lányát' (the consul's daughter + Acc) is a noun phrase (NP) headed by the noun 'lány' (daughter) – this means that this phrase is an embedded phrase which behaves as it were a single noun (in Acc).

The sentence shown in Fig. 1 could be bracketed as follows:
[[<<Gróf(NP)> Vásárhelyi(NP)> <Görögországban(NP)> <kötött ki(VV)> Clause)] és(Conj) [<titkárul(NP)> <szerződtette(VV)> <a(Art) <főkonzul(NP)> lányát(NP)> (Clause)] (Sentence)]

2.3 *Morphological analysis and disambiguation of syntactic analysis*

As explained so far, syntactic phrasing of a sentence needs morphological analysis, too, in order to identify the grammatical cases and

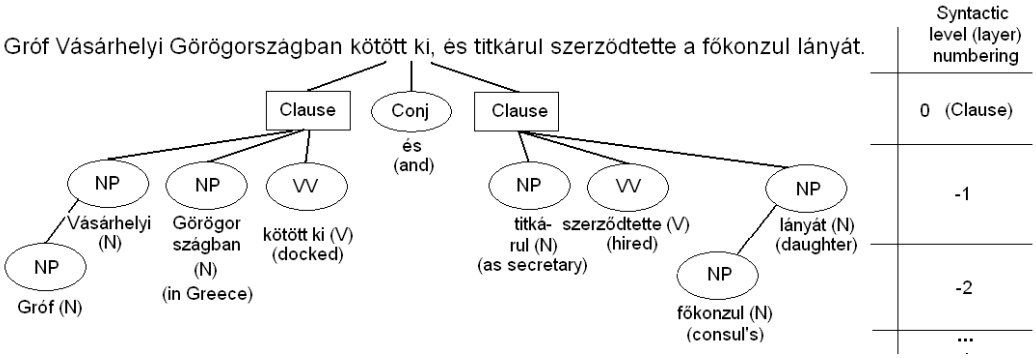


Figure 1: An example of syntactic phrase structure of the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as his secretary*)

relations in which words are used actually. This is why a morphological analyser, called Hunmorph (Trón et al., 2005), is also used from within the tool Hunpars.

The rich morphology may also lead to homonymy: words with same spelling but with different meaning, eventually also two different stems with different suffixes may result in the same word having quite different meaning and being in different cases. This causes ambiguity during the automatic syntactic analysis. Some disambiguation is performed during syntactic analysis relying on the phrase structure grammar (Babarczy et al., 2005): based on a lexicon and some rules, a part of the concurring analysis hypotheses can be ruled out. The remaining ones, however, are all kept and output by the Hunpars tool. As further automatic disambiguation is not provided by the tool, in a case of multiple hypotheses the actually correct one was selected by an expert.

3 AUTOMATIC PROSODIC SEGMENTATION OF SPEECH

3.1 *Prosodic hierarchy model*

The model of the prosodic structure used in this work relies on the *prosodic structure hypothesis* (Selkirk, 2001). This model provides a hi-

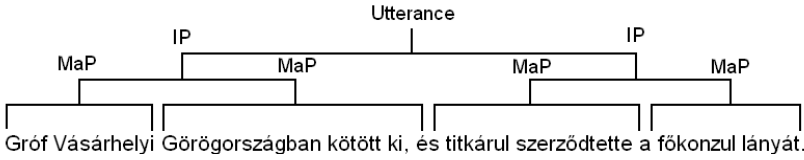


Figure 2: An example of canonical prosodic structure of a Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*"

erarchic view of the prosodic structure as follows top-down: utterances are composed of *intonational phrase*s (IP), which can be divided into *phonological phrase*s (PP). Selkirk's model differentiates between major (MaP) and minor (MiP) phonological phrases. Some studies argue that this distinction is not necessary (Ito-Mester, 2008). Indeed, the acoustic-phonetic realizations of major and minor phonological phrases seem to be very close to each other (at least for Japanese (Ito-Mester, 2008) and for Hungarian, as this issue can be language-dependent. This suggests creating a sort of recursion in the language, i.e. there is no significant difference between major and minor phonological phrases, but rather a general phonological phrase layer exists, which can embed further other phonological phrases and creates sublayers within the phonological phrase layer of the prosodic hierarchy model. However, in this work, phonological phrases are regarded as being identical with minor phonological phrases unless explicitly stated otherwise. This prosodic structure is often represented as a tree or bracketing of the utterance. An example is given in Fig. 2 for a Hungarian sentence (supposing the speaker uses the canonical prosodic patterns when uttering the sentence): "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as a secretary*). The canonical prosodic structure of this sentence could be written bracketed as: *[[<Gróf Vásárhelyi> <<Görögországban> <kötött ki és>>][<<titkárul> <szerződtette a>> <főkonzul lányát>]]*.

The prosodic hierarchy could be further refined, e.g. phonological phrases are composed of *phonological word*s, called sometimes prosodic words and so down to the syllable level, but units inferior to (minor) phonological phrases are beyond our interest in the current work, as our goal is to assess the syntax based on suprasegmental prosodic features. This means that units shorter than a phonological phrase (often

containing a single prosodic word) are not regarded as suprasegmental units in speech and fall out of interest here, as segmental domain processing of speech is a common task in automatic speech recognition (even if the used features are rather spectral or spectra-derivative and not (micro)prosodic ones), whilst suprasegmental domain processing is mostly missing in these applications (Vicsi-Szaszák, 2010) – as already mentioned in the Introduction. Another reason for focusing on the suprasegmental domain when assessing prosody is that segmental level exploitation of prosody seems to be highly language-dependent, for example in tonal languages, prosody has to be integrated into phoneme- or word-based speech recognizers in the segmental domain (Chang et al., 2000), or even in the non-tonal Japanese, prosody can be exploited in mora recognition as individual words are usually characterized by specific prosodic attributes allowing the identifications of word boundaries based on prosody (Hirose et al., 2001), (Iwano, 1999). However, whilst segmental domain use of prosody is language-dependent, the role of prosody in the suprasegmental domain is more universal and allows for the evaluation of a more general framework to assess it in this domain.

In the prosodic structure, upper-level prosodic units dominate lower-level ones, that is, for example, intonational phrase dominates the underlying phonological phrases. One of the phonological phrases belonging to the same intonational phrase usually constitutes a focal – stressed or somehow highlighted – part of the intonational phrase. In present study this means that the focus is realized with a higher F0 – local F0 peak. This phonological phrase can be called the head phrase. More generally, all phonological phrases are influenced by their location and role in the intonational phrase, which means that typical prosodic patterns can be associated with each phonological phrase. This allows us to cluster and classify individual phonological phrases and create a more or less disjunct set of phonological phrases in terms of their intonational contour, strength and location of the stress or prominence they carry, etc. In other words, clustering of phonological phrase types involves implicitly effects linked to upper level (Map or IP or utterance level) constraints and hence, phonological phrase models implicitly incorporate and reflect the upper prosodic structure of the utterance to some extent. For the Hungarian language, 6 different phonological phrase types were created in (Vicsi-Szaszák, 2005).

Prosodic label	Description
co	Clause onset PP
ss	Strongly stressed PP
ms	Medium stressed PP
ce	Low clause ending PP
cr	High ending (continuation rise) PP
ls	Low-stress PP

Table 1:
Phonological phrase types for Hungarian following Vicsi-Szaszák (2005)

They are listed in Table 1. It can be clearly seen that the distinction between them is based on the influence of higher level functions governed primarily by the intonational phrase they belong to.

The theoretic prototype of phonological phrases in Hungarian shows a smart rise of F0 at the stressed syllable, then a slightly descending contour follows. As Hungarian is a fixed-stress language (stress, if present, can almost always be found on the first syllable of the word stressed), location of the stress within the phonological phrase is not a distinctive feature.

As phonological phrases are constituents of intonational phrases, the higher level constituent influences their characteristics. Clause onset (*co*) and clause ending (*ce*) usually alter the standard phonological phrase intonational contour, so does the focus (strongly stressed phonological phrase, *ss*) and the continuation rise (*cr*). The continuation rise usually alters the subsequent phonological phrase, causing the stress to be often undetectable or turned into (low) stress (*ls*). Although Selkirk (2001) underlines that “prosody is strictly layered”, that is, higher level constituents immediately influence only the constituents located one level below, it is clear that even utterance-level constraints might have their effect on phonological phrases, as it is the case between low (ce) and high (cr) clause endings: alterations provoked by upper-level constraints propagate further down to the (minor) phonological phrase layer. These also mean that modelling done in the phonological phrase layer implicitly incorporates higher-layer information and may be used on these higher layers to perform some analysis. This hypothesis is also addressed in the paper.

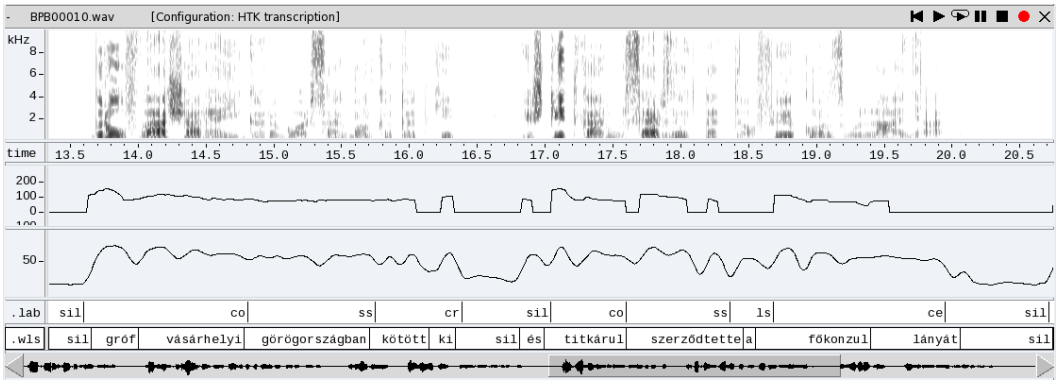


Figure 3: An example of the output of the prosodic segmenter for the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*". Spectrogram, partly interpolated F0, long time energy, prosodic segmentation and word level segmentation is also given.

3.2 *Automatic alignment of phonological phrases*

For Hungarian, an automatic phonological phrase classifier and aligner software has been made available (Vicsi-Szaszák, 2010). The parallel classification and alignment operates theoretically like a Hidden Markov Model-based automatic speech recognizer used in word or phoneme segmentation mode, but the features used are prosodic ones (see subsection 3.3) and the models are those of the phonological phrases presented in Table 1. All these mean that this tool performs automatic segmentation for phonological phrases: detects hypothetised phonological phrase boundaries and classifies the phrases. An example output of this is shown in Fig. 3. The authors underline that in this alignment approach, continuous tracking of prosody over the whole speech signal is implemented instead of looking for discrete markers, indices of breaks or tones. This provides a soft and more flexible framework, which is believed to be also closer to human perception processes.

The alignment for phonological phrases operates only at one level (or layer) in the prosodic hierarchy, that of phonological phrases. However, as phonological phrases are influenced directly by intonational-phrase-level constraints and indirectly by utterance level constraints, the prosodic structure in terms of the layering of the pro-

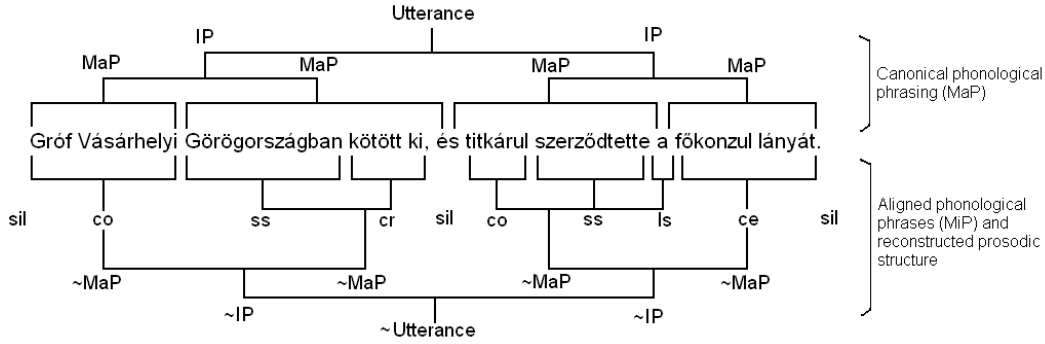


Figure 4: Reconstruction of the prosodic layering based on phonological phrase alignment for the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*".

sodic hierarchy becomes at least partly recoverable based only on the output of phonological phrase sequence aligned to an utterance. The layering of prosody will be represented in the different types of the phonological phrases, e.g. the beginning of a *clause onset (co)* phonological phrase represents the beginning of an intonational phrase and might represent the beginning of a new utterance. In a similar way, the end of a *continuation rise (cr)* phonological phrase is also the end of the embedding intonational phrase. Or, the end of a *low clause ending (ce)* phonological phrase is also the end of the embedding intonational phrase (one level higher) and the utterance itself two levels higher which embeds the intonational phrase. The prosodic segmentation of the utterance shown in Fig. 3 yields a prosodic layering (hierarchy) as shown in Fig. 4 (which is quite close to the one presented in Fig 2).

It is important to notice that the "deepness" of the prosodic segmentation based on phonological phrase alignment is speaker- and speaking style-dependent (Vicsi-Szaszák, 2010). If the speaker uses a rich prosody in her/his utterances, the deepness of the segmentation can reveal a high ratio of minor phonological phrases or even more, the boundaries of distinct words in speech (see subsection 3.6), or even separate constituents of compound words in some cases. However, if the speaker uses a "flat" prosody, the deepness of the prosodic segmentation will degrade, in extreme cases the intonational phrase cannot be divided further for phonological phrases as prosodic cues are missed or missing (deleted). The very flexible time warping capability of Hid-

den Markov models (composed of up to 11 states) explains why parts of speech of such variable length can be processed using the same phonological phrase model set and approach, i.e. an utterance composed of 6 words corresponds to one no more dividable intonational phrase in an utterance, whilst the same 6 words with the same meaning can be divided into 3 or more phonological phrases in another realization (utterance).

The prosodic segmentation is based on suprasegmental prosodic features (fundamental frequency and energy) described in subsection 3.3. During prosodic segmentation, a sophisticated prosodic phrase sequential model can be used, which constrains which phonological phrase can follow a given other one. This model has an identical role to a language model in speech recognition. The model incorporates the prosodic structure presented in subsection 3.1 and presumes that the utterance is composed of phonological phrases, which are influenced by higher level (intonational phrase or utterance level) constraints. The basic topology is as follows: each utterance begins with a clause onset phonological phrase (*co*) and ends with either a low clause ending phonological phrase (*ce*) or a high ending phonological phrase (*cr*). The utterance can embed several further intonational phrases starting with strongly stressed phonological phrases (*ss*) and ending with continuation rise (*cr*), if they are not utterance-final IPs/PPs. Stressed phonological phrases (*ss* and *ms*) are allowed to appear anywhere within the utterance. The *ss* symbol refers to a stronger accent expected to be placed at the focus of the utterance. The low-stress phonological phrase (*ls*) is allowed optionally, but only immediately after a high ending (continuation rise, *cr*). Between all utterances, a silence (*sil*) is supposed. However, this relatively strict utterance model is supposed to fit read or moderately spontaneous speech. If speech is spontaneous, a better choice can be using an utterance model which simply allows every phonological phrase entity to be aligned with no respect to its context (Szaszák-Nagy-Beke, 2011).

3.3 *Acoustic-prosodic pre-processing*

The acoustic-prosodic features used in phonological phrase models of the prosodic segmenter rely on fundamental frequency and energy. Fundamental frequency (F0) is extracted by ESPS method using a 25 ms long window. Intensity is computed with a window of 150 ms.

The frame rate for both variables is set to 10 ms. The obtained F0 contour is first filtered with an anti-octave jump tool in order to eliminate or at least reduce pitch tracking errors. This is followed by a smoothing with a 5 point mean filter. In order to ensure a relatively continuous F0 contour, a linear interpolation is carried out in logarithmic domain. However, the interpolation is omitted for voiceless sections which are longer than 150 ms or for sections where the F0 difference between the two neighbouring voiced parts shows a rise reaching at least 110% after an unvoiced part. Delta and acceleration coefficients are also appended to both F0 and intensity streams.

3.4 *Training of the prosodic segmenter*

According to Vicsi-Szaszák (2010), the training of the acoustic-prosodic models of the prosodic segmenter was performed on a part of the Hungarian database BABEL (Roach et al., 1996), hand-labelled initially for phonological phrases based primarily on the F0 contour, but also on the annotators' perception of phonological-phrase-initial stress after listening to the utterance. 1600 utterances from 32 speakers were used to train 11 state left-to-right Hidden Markov Models for each phonological phrase + silence presented in Table 1. The reason for training 11 state models (iteratively optimized during validation) is the suprasegmental nature of prosody: phonological phrases usually correspond to longer sections of speech compared, for example, to phoneme models in automatic speech recognition, which are 1- to 5-state long, but most often 3-state long).

3.5 *Initial testing of the prosodic segmenter*

Initial testing of the prosodic segmenter was carried out using 10-fold cross-validation. This means that after randomly ordering the utterances, they were divided into 10 equal subsets (160 utterances each). Training and testing was then performed 10 times by using each subset as a test set and the remaining 9 as the train set. In 10-fold cross-validation each utterance is tested in one of the cycles, but it is guaranteed that once an utterance is placed into the test set, it is excluded from the train set.

For utterances under test, a phonological phrase alignment was generated which was compared to its hand-labelling used as reference. Once these alignments were available for all utterances, four

performance indicators were measured: the recall and precision of the phonological phrase boundary recovery, the average time deviation between detected and reference phonological phrase boundaries and the accuracy of the classification regarding the type of phonological phrases.

The recall is measured with the following formula:

$$\text{Recall} = \frac{tp}{tp + fn}, \tag{1}$$

where tp stands for true positives, that is, the number of phonological phrase boundaries correctly found within 150 ms of the original one in the reference; fn stands for false negatives, that is, the number of missed phonological phrase boundaries (present in reference but not detected).

Precision is measured as:

$$\text{Precision} = \frac{tp}{tp + fp}, \tag{2}$$

where fp stands for false positives: phonological phrase boundaries detected where they should not be according to the reference, or more than 150 ms apart from reference phonological phrase boundary.

The recall of phonological phrase alignment-based prosodic segmentation was 82.1%, the precision was 77.7%.

The average time deviation (σ_t) of segmentation for phonological phrases was measured for true positives as:

$$\sigma_t = \frac{1}{tp} \sum_{i=1}^{tp} \left| t_i - t_i^{ref} \right|, \tag{3}$$

where tp stands again for the number of phonological phrase boundaries correctly found within 150 ms vicinity of the reference boundary. t_i is the detection time of the i^{th} phonological phrase boundary, t_i^{ref} is the location of the corresponding reference boundary. For the above tests, average time deviation was found to be: $\sigma_t = 50.4$ ms.

Finally, classification accuracy is measured as the ratio of correctly classified phonological phrase boundaries (tp_{cc}) versus all true positive phonological phrase boundaries (tp):

$$\text{Acc} = \frac{tp_{cc}}{tp}. \tag{4}$$

Classification accuracy was found to equal overall 73.1%.

3.6 *Prosodic segmentation vs. word boundaries*

Vicsi and Szaszák used a similar prosodic segmentation for phonological phrases to partially recover word boundaries in Hungarian and Finnish languages (Vicsi-Szaszák, 2010), (Vicsi-Szaszák, 2005). Of course not all phonological phrase boundaries coincide with word boundaries, the authors also underline that for Hungarian, a word boundary detector in the strict sense cannot be implemented in contrast to the mentioned Japanese (Hirose et al., 2001). However, they trained the prosodic-acoustic models of phonological phrases on samples in which phonological phrase boundaries coincided with word boundaries. Highly relying on the first syllable fixed stress of Hungarian, word boundaries were predicted in the vicinity of phonological phrase boundaries. Analysis of word boundary detection rates based on phonological phrase alignment showed 77.3% precision and 57.2% recall rate for Hungarian (on BABEL speech database), 69.2% precision and 76.8% recall rate for Finnish allowing a maximum of ± 100–150 ms deviation between phonological phrase and word boundary markers (Vicsi-Szaszák, 2005). The goal of the experiments described in present paper can be related to this issue, namely, to prove or to disclaim the conjecture that the detected word boundaries correlate well with syntactic phrase boundaries, while missed word boundaries are more likely to be embedded within a syntactic phrase, and therefore tend to form a union both prosodically and syntactically.

4 ANALYSING THE PROSODY-TO-SYNTAX MAPPING

The main goal of the paper is to present a detailed analysis regarding the prosody-to-syntax automatic mapping possibilities in spoken language. This implies the comparison between the prosodic and syntactic structures, obtained based on analyses presented so far both for prosody and syntax. The syntactic phrasing will be used as reference, and hence – although it was primarily obtained in automatic way – it has to be checked and disambiguated by human experts. The automatically obtained prosodic phrasing on the other hand is left intact as it is produced by the prosodic segmenter tool. The reason for this is that this approach will permit to evaluate the usability of the pro-

posed algorithm in real conditions, where the automatically obtained prosodic phrasing may contain errors.

4.1 *Material and method*

The material used for current experiments was taken from the BABEL Hungarian language speech database (Roach et al., 1996). BABEL is a read speech database, involving 60 non-trained speakers' data. The speech material covers paragraphs composed of at least 6 sentences (contextually linked), numbers, isolated digits, and CVC items. A subcorpus taken from the paragraphs was used, containing 155 different sentences uttered by a total of 60 speakers. Most of the sentences occurred at least two times, hence a total set of 330 sentences was used.

The utterances were segmented on word level, obtained by performing automatic forced alignment on word-level transcriptions with a Hungarian language ASR. Indeed, an automatic phoneme segmentation was performed, which was traced back to word-level alignment. This means that time positions of word boundaries were known. This will be necessary for temporal word boundary information, as syntactic phrase boundaries themselves are located always on word boundaries.

In order to obtain the syntactic analysis, sentences (transcriptions of the speech utterances) were fed into the Hunpars syntactic analyser. Where the syntactic analyser yielded unresolved ambiguity, a human expert intervened in order to leave one and only one syntactic parsing for every sentence contained in the speech utterances (as minimal pair sentences were not included in the material, there was always only one canonically correct syntactic parsing candidate for each sentence).

In parallel, speech utterances were fed into the prosodic segmenter tool too. This produced the phonological phrase alignment of the utterances. As the prosodic segmenter tool was also trained on the BABEL database (and a testing of 10-fold cross-validation was also done for it as described in subsection 3.5), special attention was paid to ensure that the utterance currently under analysis is excluded from the training set of the prosodic segmenter.

Hereafter, the correspondence of the automatically detected phonological phrase boundaries and syntactic phrase boundaries was investigated. Correspondence was assessed separately on each syntactic level (layer) of the syntactic hierarchy, to a depth of 5 levels (top-

down: 0, −1, −2, −3, −4): sentences are divided into clauses (level 0). Clauses consist of first level syntactic phrases (level −1), which can contain daughter phrases (level −2) and so on down to level −4. While levels 0, −1 and −2 are quite common and occur in most of the sentences, deeper embedding (level −3 and especially level −4) is quite rare. The numbering of the syntactic layers is included in Fig. 1 for evidence.

The main interest is to see whether syntactic phrase boundaries can be detected based on phonological phrase alignment, and whether the syntactic hierarchy (layering) can be reconstructed based on the aligned phonological phrases (or prosodic layering, as the differentiation among phonological phrases allow for the reconstruction of the latter, as explained in subsection 3.2 and in Fig. 4).

Syntactic and phonological phrase boundaries were considered to meet if they occurred within 150 ms time interval. This value was chosen based on the following considerations:

- the time interval should allow some deviation in a range of a length of a demi-syllable, because reference word boundaries (necessary for the identification of the onset and ending times of the syntactic phrases) are segmented automatically, and
- the prosodic segmenter also displays some uncertainty (for example, if an utterance ends with an unvoiced sound, it is often inevitably chopped).
- phonological phrases aligned by the prosodic segmenter are much longer than 150 ms (average phonological phrase length is 618 ms for the test corpus with a standard deviation of 211 ms).

Syntactic phrase (XP) onsets were always aligned to phonological phrase (PP) onsets, syntactic phrase endings were always aligned to phonological phrase endings.

4.2 *Recovering syntactic phrase boundaries*

In the first experiment, phonological phrase segmentation is used to recover syntactic phrases automatically. The performance is evaluated using the *recall* measure defined in equation (1), but now tp stands for correctly recovered syntactic phrase boundaries (true positives) and fn stands for missed syntactic boundaries (false negatives). The type of the aligned phonological phrase is irrelevant in this exper-

Table 2: Recall of syntactic phrase boundaries by phonological phrase boundaries summarized for all phonological phrase types. 1B/L = one (highest level) syntactic boundary kept per word; MB/W = multiple syntactic boundaries allowed per word

Syntactic	Onset		Ending		# of occ.
Level	**1B/W**	MB/W	**1B/W**	MB/W	(MB/W)
0	**0.85**	0.85	**0.79**	0.79	3124
−1	**0.45**	0.70	**0.48**	0.68	10339
−2	**0.42**	0.70	**0.48**	0.69	5763
−3	**0.44**	0.74	**0.45**	0.65	814
−4	**0.48**	0.70	**0.50**	0.67	187
All	**0.54**	0.72	**0.55**	0.69	20227

imental setup, currently the only interest is to see what portion of syntactic phrase boundaries are detectable on the different syntactic layers, based on the phonological phrase alignment.

Results are shown in Table 2 separated for phrase onsets and phrase endings. As multiple-level syntactic embeddings are possible, several syntactic boundaries can occur at the same place. In one scenario, only the highest-level syntactic boundary is counted in case of multiple level occurrences (referred to as 1B/W), while in the other one, all different level syntactic boundaries are counted (MB/W). This means that a word preceded by level 0, −1 and −2 syntactic boundaries yields one 0 level hit (true positive) in 1B/W if detected, but gives a total of 3 hits, one for each level 0, −1 and −2 in MB/W if detected (and, of course, gives false negatives for all the 3 levels if remains undetected).

Average recall rate was 71% (in MB/W) or 55% (in 1B/W), which is considerably higher in the case of clauses: 85% (onsets) and 79% (endings). A total of around 70% of the syntactic phrase boundaries can be detected on each syntactic layer. Deeper syntactic embedding did not seem to degrade detection rates. For statistical evidence Kruskal-Wallis tests were performed on the obtained data. These also confirm that phonological and syntactic phrases are correlated ($\chi^2 = 6430.606; p < 0.000$).

Pairing up the corresponding syntactic phrase onset and syntactic phrase ending boundaries on each level and comparing recall rates by Mann-Whitney and Wilcoxon W tests show that on clause level, onsets are significantly better detected ($Z = -7.807; p < 0.000$). However,

on levels -1 and -2, there is no significant difference in recall rates counted for onsets and endings (level 1: $Z = -0.407, p > 0.1$; level 2: $Z = -0.016; p > 0.1$). There were no significant differences either on levels -3 and -4.

Non-clause (< 0) syntactic levels do not yield different recall rates (either in 1B/W or in MB/W settings), this means that lower level syntactic phrases are not less intensively marked by prosody: clauses can be identified by a higher recall rate, but there is no significant difference between syntactic phrases depending on the syntactic layer ($\chi^2 = 0.224; p > 0.1$). Each syntactic phrase seems to behave as an independent entity in terms of detectable prosodic features, independently of its position in the syntactic hierarchy. These findings may also support theories supposing a recursive nature of speech prosody (cf. Wagner, 2005).

4.3 *Reliability of the syntactic phrase recovery*

In the next step, the reliability of the segmentation was analysed, separated for all phonological phrases (except for silence (*sil*)). The reliability of the phonological phrase alignment (i.e if a phrase boundary is detected based on prosody, to what extent it is sure that there is a real syntactic boundary there or that the hit is not a false one) is measured with precision according to equation (2), but now tp stands for the number of phonological phrase boundaries which coincide with syntactic phrase boundaries, and fp stands for inserted phonological phrase boundaries which do not coincide with syntactic phrase boundaries within 150 ms (false positives). Precision measures are shown in Fig. 5 for phrase onsets and endings, separated for each phonological phrase type used. As it can be seen, the onset of a *co* phrase yields a good precision rate for syntactic phrase onsets, in parallel with the hypothesis that the beginning of a *ce* phrase should refer to a level 0 syntactic phrase, which is prosodically better marked than deeper syntactic phrases. The ending of the *ce* phrase is associated more often with deeper and hence less prosodically marked syntactic phrases (see later Table 3). Phonological phrase endings of *ce* and *cr* phrases give high precision for syntactic phrase endings: again, this can refer to corresponding level 0 syntactic phrase endings which are prosodically better marked. (see later Table 4) These hypotheses are addressed in the next subsection (subsection 4.4).

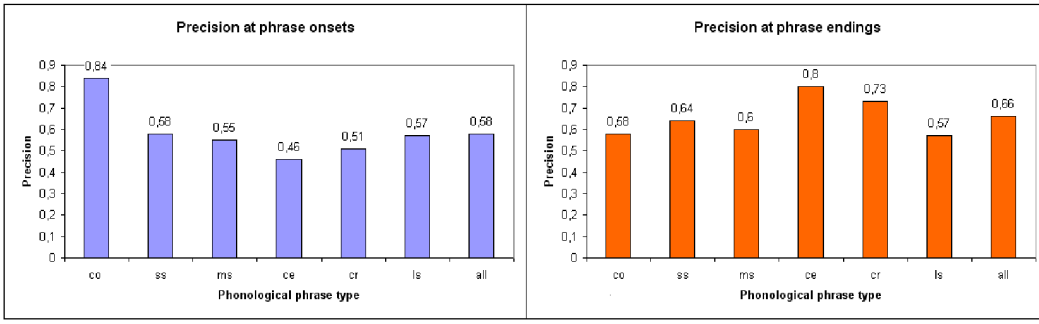


Figure 5: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right)

4.4 *Towards a reconstruction of syntactic layering*

As presented in subsection 3.2, the prosodic layering can be – at least partially – reconstructed based on the type of the phonological phrases. The next analysed point is whether there can be found some interconnection between the type of phonological phrase and the position in the hierarchy of the syntactic phrase they refer to. This could also explain differences in precision seen in Fig. 5 and justify the hypotheses raised. This would mean that not only the syntactic phrase boundaries, but also the syntactic structure in terms of its layering may become recoverable based on phonological phrase alignment.

The distribution of the aligned phonological phrases was hence examined on each syntactic layer, separately, in order to see whether some types of phonological phrases can be associated with specific syntactic layers or not. Tables 3 (for phrase onsets) and 4 (for phrase endings) show relative frequencies of the layer position of the recovered syntactic phrase (to which layer it belongs to in the syntactic hierarchy) depending on the type of the phonological phrase.

Based on the results in Table 3, a detected *co* type phonological phrase onset corresponds to a clause onset with 86% relative frequency. This means that this type of phonological phrase is a good indicator of a clause onset. Level −1 syntactic phrase onsets are well predictable if the phonological phrase type is *ss*, *ms*, *ce*, or, to a lesser extent, *cr*. Phonological phrase type *ls* onset is ambiguous, it can sign both a clause onset (50% rel. frequency) and a first level syntactic phrase onset (41%). Down from syntactic level −2, all phonologi-

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	**0.86**	0.07	0.04	0.02	1736
ss	0.12	**0.78**	0.07	0.02	2517
ms	0.09	**0.83**	0.06	0.01	1399
ce	0.14	**0.80**	0.04	0.02	2094
cr	**0.22**	**0.72**	0.04	0.01	1326
ls	**0.50**	**0.41**	0.07	0.02	1467
all	0.36	0.56	0.05	0.02	10539

Table 3: Distribution of syntactic phrase (XP) levels (or layers) based on phonological phrase type (phonological phrase onsets compared to syntactic phrase onsets)

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.05	**0.74**	0.11	0.08	1736
ss	0.09	**0.68**	**0.20**	0.03	2517
ms	0.08	**0.68**	**0.18**	0.04	1399
ce	**0.83**	0.11	0.04	0.02	2094
cr	**0.60**	**0.28**	0.09	0.03	1326
ls	0.13	**0.64**	**0.17**	0.06	1467
all	0.34	0.49	0.13	0.04	10539

Table 4: Distribution of syntactic phrase (XP) levels (or layers) based on phonological phrase type (phonological phrase endings compared to syntactic phrase endings)

cal phrase types are distributed uniformly, the aligned phonological phrase type cannot be used to predict syntactic level. Results prove that intonational phrases and clauses are very closely related, and that clauses can be automatically well separated from lower-level syntactic phrases. This means that two layers of the syntactic hierarchy can be accurately recovered: level 0 and lower levels, which cannot be further separated (but levels under level −1 occur much more rarely then level −1 phrases and hence, the major skeleton (the top) of the syntactic structure can be recoverable).

The detected *ce* phonological phrase ending mostly corresponds to a clause ending, this is approved by the 83% frequency (Table 4). The ending of a phonological phrase of type *cr* signs often a clause ending (60%), although it can also correspond to a level −1 syntactic phrase ending with a relatively high frequency (28%). Ending of phonological

phrases of types *co* predict a level −1 syntactic phrase ending with 74% frequency, endings of phonological phrases *ss*, *ms* and *ls* can refer to level −1 and −2 syntactic phrase endings. Levels −1 and lower levels cannot be separated further based on the comparison of endings of phonological phrases and syntactic phrases.

4.5 *Head classification of the syntactic phrase*

Relations between the types of phonological (co, ss, ms ce, cr, ls) and syntactic phrases (NP, AdjP, AdvP, NumP, VV, VV-Inf, PostpP) were also investigated. It was found that there is no significant difference in the phonological phrase type depending on the type of syntactic phrase in the Hungarian language ($\chi^2 = 0.349; p > 0.1$). This result is not surprising, especially with regard to the relatively free word order of Hungarian. In other languages, where the position of syntactic constituents (words or phrases) refers to grammatical relations, syntactic phrase classification (based on the head) might be possible, as clause onsets and endings are known, however, this issue would need further experimental examination and evaluation. For morphologically rich languages characterized by a more free word ordering, morphological analysis seems to be unavoidable for such purposes. In speech-based systems, this involves the use of automatic speech recognition, the output of which could be segmented to phonological phrases, and then fed into a syntactic parser and morphological analyser.

4.6 *Robust intonational phrase – clause recovery*

Precision and recall of phonological phrase segmentation were also analysed with a reduced phonological phrase set: *ms* and *ls* phonological phrase types were discarded during the phonological phrase alignment, as *ss* was expected to replace *ms*. Phonological phrase type *ls* was discarded because it yielded ambiguous results in syntactic phrase onset detection regarding the identification of the syntactic level. Results for phonological phrase/syntactic phrase onsets show (see table 5 for phrase onsets and table 6 for phrase endings) significantly higher precision (overall 64% for onsets, 65% for endings, see Fig. 6) and lower recall (overall 48% at onsets, 47% at endings, 1B/W) rates, while precision of phonological phrase/syntactic phrase ending detection is not significantly different, but recall rates are worse. The lower recall can be explained by the fact that phonological phrases with less charac-

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.88	0.07	0.02	0.02	1835
ss	0.13	0.77	0.07	0.02	3455
ce	0.26	0.67	0.04	0.02	1914
cr	0.37	0.58	0.04	0.01	1782
all	0.42	0.51	0.05	0.02	8986
Recall	0.80	0.39	0.34	0.37	

Table 5:
Distribution of syntactic phrase (XP) layers based on phonological phrase type with reduced phonological phrase set (onsets compared to onsets), 1B/W recall is also shown in the last line

Phonological phrase type	Distribution of XP levels 0	−1	−2	−3	# of occ.
co	0.03	0.43	0.07	0.04	1835
ss	0.05	0.45	0.12	0.02	3455
ce	0.50	0.12	0.03	0.01	1914
cr	0.41	0.19	0.06	0.03	1782
all	0.21	0.32	0.08	0.02	8986
Recall	0.67	0.41	0.39	0.39	

Table 6:
Distribution of syntactic phrase (XP) layers based on phonological phrase type with reduced phonological phrase set (endings compared to endings), 1B/W recall is also shown in the last line

teristic stress (*ms* and *ls*) are sometimes identified as *ss* but may also remain undetected (phonological phrase *ss* cannot replace all of their occurrences). Interpreting these in a prosodic hierarchy point of view, this approach seems to operate on major phonological phrase layer and not on the minor one. As it allows for more precise clause onset detection (see Table 5), it can be used individually or combined to the minor phonological phrase alignment based recovery approach (subsection 4.4) if higher precision is required in the reconstruction of the top layer of the prosodic hierarchy.

5 CONCLUSIONS

In the paper, automatic recovery of the syntactic structure was addressed based on prosody. The output of a phonological phrase level segmentation tool was used to predict syntactic phrase boundaries. Up to 85% of the clause boundaries and about 50% of further non-

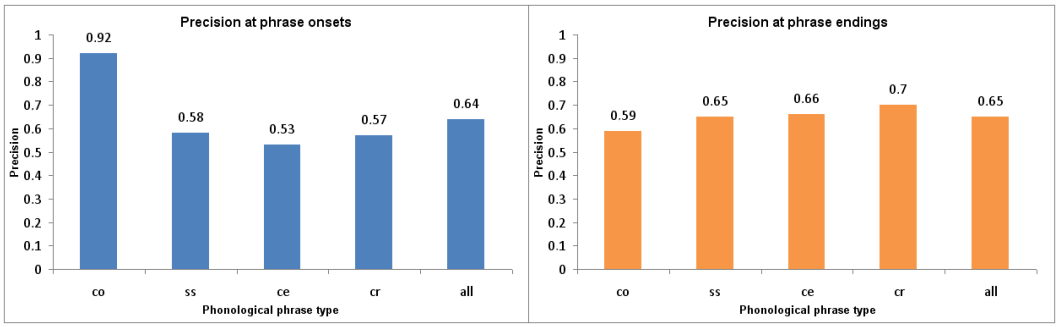


Figure 6: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right), using the reduced phonological phrase set

coinciding lower-level syntactic phrase boundaries could be automatically recalled. Precision of clause boundary detection (i.e. when intonational phrase boundaries met clause boundaries) was 84% (even 92% with a reduced phonological phrase model set), precision of lower level syntactic boundary detection (i.e. syntactic phrase boundary met by phonological phrase boundary) ranged between 46% and 58%, allowing at most 150 ms deviation between the phonological and the syntactic boundary markers. Clause level and underlying syntactic phrase level could be well separated based on the type of the aligned phonological phrase.

No relation was found between the type of the syntactic phrase and the type of the phonological phrase. This is not surprising, as the investigated language was the Hungarian, characterized by free word order. Based on the results presented, prosody seems to have a synchronizing and signalling function in terms of identification of the underlying syntactic units, but does not seem to reflect the finer relations among these units in lower syntactic layers. These results also raise some evidence of the recursive nature of speech prosody: syntactic boundaries are well signalled by prosody irrespective of the syntactic layer by same recall rates for each layer with no significant difference among them), but after a point in the hierarchy (layers down from layer -1 in the syntactic structure and layers of minor phonological phrases in the prosodic structure), layering information disappears from prosody, but layer boundaries remain detectable with the same

accuracy. This can suggest a hypothesis that at this point semantics takes over the layering role from prosody, however, this issue needs further investigation.

Although the results shown in the paper were obtained for the Hungarian language, no language-specific knowledge was used during the experiments, per se, the syntactic analyser and prosodic segmenter modules are language specific. The prosodic segmenter module, on the other hand, has already been successfully used for Finnish and German languages (Vicsi-Szaszák, 2010). The presented results can highly contribute to support automatic speech understanding. Possible application areas of the results can be speech segmentation based on prosody for supporting meaning extraction, surface syntactic structure analysis based on speech, support for text-based syntactic analysis, topic-comment separation, keyword spotting where the keyword is stressed, etc.

ACKNOWLEDGEMENTS

Authors express their gratitude to Alexandra Markó, ELTE University, Budapest, Hungary and to former MSc. student Katalin Nagy at the Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics for their help in the experiments presented in this paper.

Authors would like to thank the CESAR (http://cesar-project.net/) project, funded under the ICT-PSP (Grant Agreement No. 271022), a partner of META-NET (http://meta-net.eu), for its support for the work done on the BABEL corpus.

REFERENCES

A. Babarczy, G. Bálint, G. Hamp, A. Rung (2005), Hunpars: a rule-based sentence parser for Hungarian, *Proc. of the 6th International Symposium on Computational Intelligence*, Budapest, Hungary.

A. Batliner , B. Möbius , G. Möhler , A. Schweitzer and E. Nöth (2006), Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground, *Proc. Eurospeech 2001, Vol. 4.*, Aalborg, Denmark, pp. 2285–2288.

S. Becker, M. Schröder, W.J Barry (2006), Rule-based Prosody Prediction for German Text-to-Speech Synthesis, *Speech prosody*, Dresden, Germany, p. 31

J. BUTZBERGER, H. MURVEIT, E. SHRIBERG and P. PRICE (1992), Spontaneous speech effects in large vocabulary speech recognition applications, *Proceedings of the 1992 DARPA Speech and Natural Language Workshop*, pp. 339–343.

E. CHANG, J.-L. ZHOU, S. DI, C. HUANG and K.-F. LEE (2000), Large vocabulary Mandarin speech recognition with different approaches in modeling tones, *International Conference on Spoken Language Processing*.

A. CHRISTOPHE, S. PEPERKAMP, C. PALLIER, E. BLOCK, amd J. MEHLER (2004), Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language*, Vol. 51, pp. 523–547.

F. GALLWITZ, H. NIEMANN, E. NÖTH, W. WARNKE (2002), Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36. pp. 81–95.

G. GAZDAR, E.H. KLEIN, G.K. PULLUM and I.A. SAG (1985), *Generalized Phrase Structure Grammar*, Oxford: Blackwell, and Cambridge, MA: Harvard University Press.

K. HIROSE, N. MINEMATSU, Y. HASHIMOTO and K. IWANO (2001), Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of Fundamental Frequency Contours, *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ, USA, pp.61–66.

J. HIRSCHBERG (1993), Pitch Accent in Contex: Predicting Intonation and Prominence from Text, *Arftificial Intelligence*, Vol. 63., No. 1–2.

J. ITO and A. MESTER (2008), *Rhythmic and interface categories in prosody* Ms., UC Santa Cruz. Presented at PRIG (Prosody Interest Group), UCSC.

K. IWANO (1999), Prosodic Word Boundary Detection Using Mora Transition Modeling of Fundamental Frequency Contours – Speaker Independent Experiments. *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 99)*, Budapest, Hungary, vol.1, pp.231–234.

E.M. KAISSE (1985), Connected Speech: The Interaction of Syntax and Phonology, *Academic Press*, San Diego.

K. É KISS (2002), The syntax of Hungarian. *Cambridge University Press*, UK.

I. KOUTNY, G. OLASZY, P. OLASZI (2000), Prosody prediction from text in Hungarian and its realisation in TTS conversion, *International Journal of Speech Technology*, Vol. 3–4, pp. 187–200.

X. LI, Y. YANG, Y. LU (2010), How and when prosodic boundaries influence syntactic parsing under different discourse contexts: An ERP study *Biological Psychology*, Volume 83, Issue 3, March 2010, pp. 250–259.

E. NÖTH, A. BATLINER, A. KIESSLING, R. KOMPE, and H. NIEMANN (2000), Verbmobil: the use of prosody in the linguistic components of a speech understanding system, *IEEE Trans*, ASSP, Vol. 8, pp. 519–532.

C. POLLARD, I.A. SAG (1994), Head-Driven Phrase Structure Grammar. *University of Chicago Press*.

P.J. PRICE, M. OSTENDORF,S. SHATTUCK-HUFNAGEL, C. FONG (1991), The use of prosody for syntactic disambiguation, *Journal of the Acoustical Society of America* Vol. 90 No. 6, pp. 2956–2970.

P. ROACH et al. (1996), BABEL: An Eastern European multi-language database, *Proc. of the* 4^{th} *International Conference on Speech and Language Processing*, Philadelphia, USA, Vol 3. pp. 1892–1893.

E. SELKIRK (2001), *The Syntax-Phonology Interface*, in N.J. SMELSER and P.B. BALTES (Eds), International Encyclopaedia of the Social and Behavioural Sciences, Oxford: Pergamon, pp. 15407–15412.

E. SHRIBERG, A. STOLCKE, Direct modeling of prosody: An overview of applications in automatic speech processing, Proc. ISCA International Conference on Speech Prosody, 2004.

E. SHRIBERG, A. STOLCKE, D. HAKKANI-TÜR, G. TÜR (2000), Prosody-Based Automatic Segmentation of Speech into Sentences and Topics, Speech Communication 32(1–2), 127–154.

K. SILVERMAN (1993), On costumizing prosody in speech synthesis: names and addresses as a case in point, *Proc. ARPA Workshop on Human Language Technology*, pp. 317–322.

K.N. STRELNIKOV, V.A. VOROBYEV, T.V,CHERNIGOVSKAYA, S.V. MEDVEDEV (2006), Prosodic clues to syntactic processing – a PET and ERP study, *NeuroImage* Volume 29, Issue 4, pp. 1127–1134.

M. SZARVAS, T. FEGYÓ, P. MIHAJLIK, P. TATAI (2000), Automatic Recognition of Hungarian: Theory and Practice. International Journal of Speech Technology 3:(3–4) pp. 237–251.

Gy. SZASZÁK, K. NAGY and A. BEKE (2011), Analysing the correspondence between automatic prosodic segmentation and syntactic structure, Proc of Interspeech 2011, Florence, Italy, pp. 1057–1061.

V. TRÓN, L. NÉMETH , P. HALÁCSY, A. KORNAI, Gy. GYEPESI, D. VARGA (2005), Hunmorph: Open source word analysis. Proceedings of the ACL 2005 Workshop on Software, Ann Arbor, MI, pp. 77–85.

N.M. VEILLEUX, M. OSTENDORF (1993), Prosody/parse scoring and its application in ATIS. Proc. ARPA Human Language Technology Workshop ’93. pp 335–40.

K. VICSI and Gy. SZASZÁK (2005), Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features, *International Journal of Speech Technology*, Vol. 8 No. 4, pp. 363–370.

K. VICSI, Gy. SZASZÁK (2005), Using prosody to improve automatic speech recognition, *Speech Communication* Vol. 52, No. 5, pp. 413–426.

M. WAGNER (2005), *Prosody and recursion*, Ph.D. dissertation, MIT.

CEntral and South-east europeAn Resources (CIP-ICT-PSP.2010.6.1)
`http://www.cesar-project.net`
Duration: 01-02-2011 – 31-01-2013

CONSORTIUM

Hungarian Academy of Sciences,
Research Institute for Linguistics, *Hungary*
Budapest University of Technology and Economics, Department of Telecommunications and Media Informatics, *Hungary*
University of Zagreb, Faculty of Humanities and Social Science, *Croatia*
Linguistic Engineering Group, Institute of Computer Science,
Polish Academy of Sciences, *Poland*
University of Łódź, *Poland*
University of Belgrade, Faculty of Mathematics, *Serbia*
Institut Mihajlo Pupin, *Serbia*
Imperial Institute for Bulgarian Language,
Bulgarian Academy of Sciences, *Bulgaria*
Institute of Linguistics, Slovak Academy of Sciences, *Slovakia*

POLISH PARTICIPANTS

Linguistic Engineering Group
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warszawa
`http://zil.ipipan.waw.pl`
Contact person: **Maciej Ogrodniczuk,** Assistant professor
Email: `maciej.ogrodniczuk@ipipan.waw.pl`
Tel.: +48 22 380 05 63

PELCRA Group, University of Łódź
al. Kościuszki 65, 90-514 Łódź
`http://pelcra.pl`
Contact person: **Piotr Pęzik,** Assistant professor
Email: `pezik@uni.lodz.pl`
Tel.: +48 42 665 52 11

Human language technologies crucially depend on language resources and tools that are usable, useful and available. However, even where language resources and respective tools are available they have been developed mostly in a sporadic manner, in response to specific project needs, with relatively little regard to their long-term sustainability, intellectual property rights status, interoperability, reusability in different contexts as well as to their potential deployment in multilingual applications.

The CESAR project (Central and South-East European Resources), a part of META-NET initiative, intends to address this issue by enhancing, upgrading, standardizing and cross-linking a wide variety of language resources and tools and making them available, thus contributing to an open linguistic infrastructure. The comprehensive set of language resources and tools covers the Hungarian, Polish, Croatian, Serbian, Bulgarian and Slovak languages. The resources include interoperable mono- and multilingual speech databases, corpora, dictionaries and wordnets and relevant language technology processing tools such as tokenizers, lemmatizers, taggers and parsers. They are being made available at partners' sites and their metadata descriptions contributed to the META-SHARE digital exchange facility.

In addition to the technical work required, great effort is dedicated to ensure sustainability through mobilizing the LT community, raising awareness of the fundamental role of language resources among the R&D policy makers, the media and the general public.