Designing a scientific journal on the example of Journal of Language Modelling

Adam Twardoch
Marcin Woliński

Publishing a scientific journal

Editorial Board
peer review

Publisher
proofreading
typesetting
copyediting
printing
distribution

Publishing a scientific journal

Editorial Board
peer review

Publisher
~~proofreading~~
typesetting
copyediting
printing
distribution

Publishing a scientific journal

Editorial Board
peer review

Publisher
~~proofreading~~
~~typesetting~~
copyediting
printing
distribution

Publishing a scientific journal

Editorial Board
peer review

Publisher
~~proofreading~~
~~typesetting~~
~~copyediting~~
printing
distribution

Publishing a scientific journal

Editorial Board
peer review

Publisher
~~proofreading~~
~~typesetting~~
~~copyediting~~
~~printing~~
~~distribution~~

Open access (OA) is the practice of providing unrestricted access via the Internet to peer-reviewed scholarly journal articles. OA is also increasingly being provided to theses, scholarly monographs and book chapters.

Open access comes in two degrees: Gratis OA is no-cost online access, while Libre OA is Gratis OA plus some additional usage rights.

[Wikipedia]

Journal of Language Modelling is a free (for readers and authors alike) open-access peer-reviewed journal aiming to bridge the gap between theoretical linguistics and natural language processing.

Journal of Language Modelling

uses double-blind peer reviews
available on the Internet
with the print on demand option

papers appear on the website as soon
as they have been accepted
twice a year an issue gets “closed”

Open Journal Systems
`http://pkp.sfu.ca/ojs/`

JLM's production process

several rounds of reviews possible

author submits TeX sources

copy-editing in the TeX sources

typesetting

proofreading

Copy-editing

A number of applications in automatic speech understanding require some analysis of the content prior or parallel to speech-to-text conversion often referred to ~~often~~ as automatic speech recognition. In speech understanding, a pure transcription of the speech yielded by a speech-to-text converter (~~speach~~speech recognizer) would be insufficient, as the underlying meaning remains unextracted, uninterpreted.

JLM: change word order

```
…converter (¶speach¶speech¶ recognizer)…
```

Toposław

Journal of Language Modelling Vol 0, No 0 (1972)
ISSN 666-666-666, pages 1–13

Toposław – a lexicographic framework for multi-word units

Małgorzata Marciniak[1], Agata Savary[2,1], Piotr Sikora[1], and Marcin Woliński[1]

[1] Institute of Computer Science, Polish Academy of Sciences,
J.K. Ordona 21, 01-237 Warszawa, Poland
[2] Université François Rabelais Tours,
3, pl. Jean-Jaurès, 41029 Blois, France

Abstract

The paper presents a tool for the creation of an electronic dictionary of multi-word proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 Introduction

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

We are interested in a lexical description of multi-word names in a particular application domain: the urban transportation system in Warsaw. In Marciniak *et al.* (2009) we present a project of creating a dictionary of Polish toponyms relevant to Warsaw transportation. The project includes the development of a lexicographic framework, *Toposław*, that allows us the creation of the dictionary in an efficient and quality-ensured manner. *Toposław* cooperates with: (i) *Morfeusz SGJP* (a new version of *Morfeusz* Woliński (2006)), a morphological analyzer and

Toposław – a lexicographic framework for multi-word units

Małgorzata Marciniak[1], Agata Savary[2,1], Piotr Sikora[1], and Marcin Woliński[1]

[1] Institute of Computer Science, Polish Academy of Sciences,
J.K. Ordona 21, 01-237 Warszawa, Poland
[2] Université François Rabelais Tours,
3, pl. Jean-Jaurès, 41029 Blois, France

Abstract

The paper presents a tool for the creation of an electronic dictionary of multi-word proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper

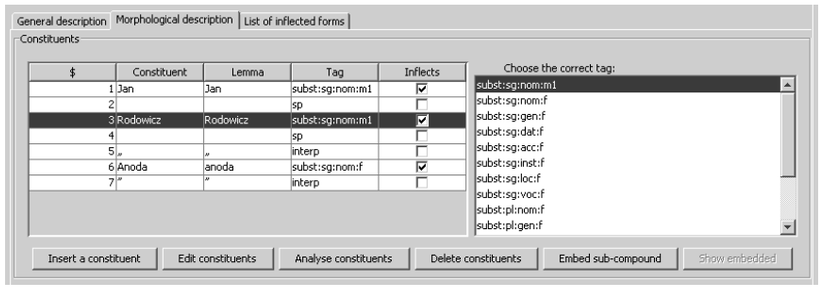


FIGURE 1: The labelled lemma for the name *Jan Rodowicz „Anoda"*

Rodowicz, J. "Anoda" Rodowicz, J. Anoda Rodowicz, "Anoda" Rodowicz, Anoda Rodowicz, Jan Rodowicz, J. Rodowicz, Rodowicz

3.1 Morphological Description of Components

Clearly, most variants of a multi-word unit can be obtained by combinations of inflected forms of their components. Therefore the morphological description of individual words, both common words and proper names, is a necessary prerequisite for our dictionary. In *Toposław* this description is ensured by *Morfeusz SGJP*, a morphological analyser and generator for Polish single words based on the data of the *Grammatical Dictionary of Polish* Saloni *et al.* (2007). For the needs of our dictionary *Morfeusz SGJP* (further called *Morfeusz*) has been extended with 1612 entries: 1005 nouns (not surprisingly last names and geographical names) and 607 adjectives (mainly adjectives derived from geographical names, e.g., *kabacki* from *Kabaty*, but also some other less frequently used adjectives like *arbuzowy*, an adjective from *watermelon*).

The *Morfeusz* analyzer provides all possible morphological interpretations of a word. This allows us to label the constituents of a compound name with their lemmas and morphological features. As shown in Fig. 1 (in the *Constituents* panel), the name is first tokenized and tokens are numbered, whereas separators (blank spaces, hyphens, etc.) and non alphabetical items (quotes, numbers, etc.) are considered as separate tokens. Then, the tokens are analyzed by *Morfeusz* and if any of them have multiple interpretations, the operator has to select the appropriate one. Here, the interpretation selected for *Rodowicz* is substantive, singular, nominative, masculine human (*subst:sg:nom:m1*).

The operator can also mark a fragment of the name as a sub-compound to be described separately. We use this mechanism for compound names of persons, which tend to occur in several urban names. For instance in Fig. 2 seven tokens of a street name have been grouped in order to create a single compound component *Jana Rodowicza "Anody"* which corresponds to the person name entry discussed above. Delimiting embedded structures within long names such as *Aleja Jana Rodowicza „Anody"* 'Jan Rodowicz "Anoda" avenue' allows for a better modularity and compactness of description.

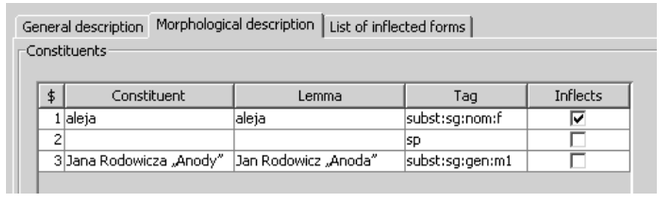


FIGURE 2: The name *Aleja Jana Rodowicza „Anody"* 'Jan Rodowicz "Anoda" avenue' after *Jana Rodowicza „Anody"* has been marked as a sub-compound

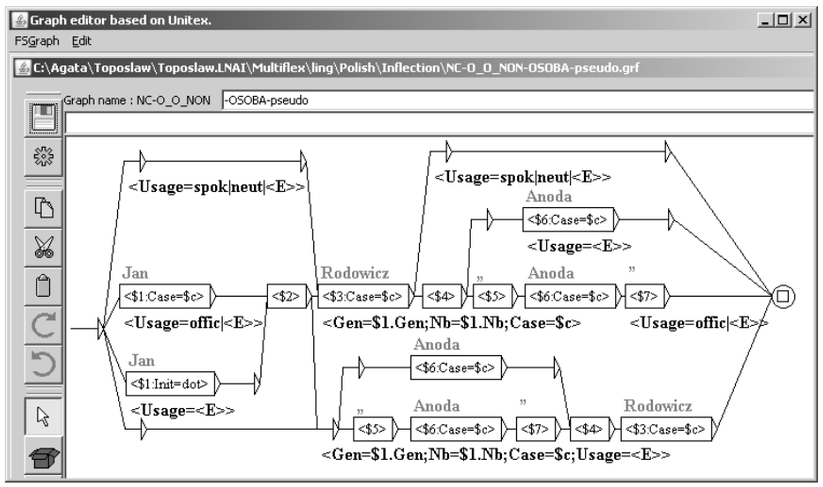


FIGURE 3: Inflection graph for the name *Jan Rodowicz „Anoda"*

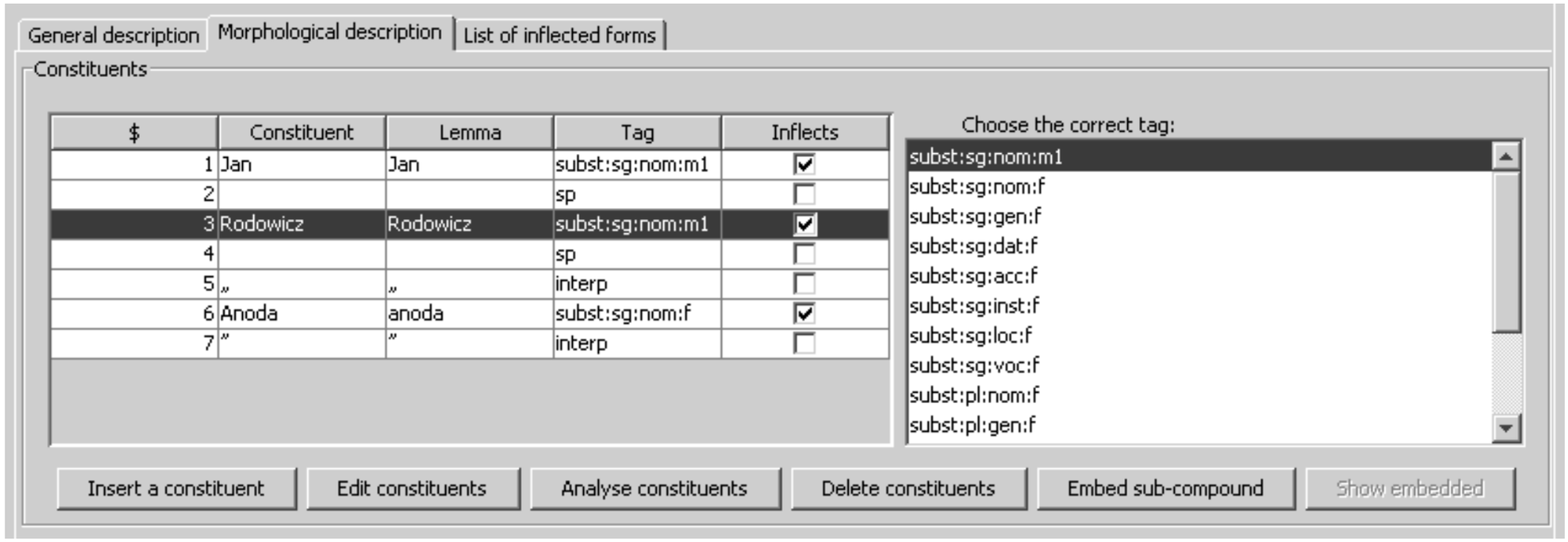


FIGURE 1: The labelled lemma for the name *Jan Rodowicz „Anoda”*

Rodowicz, J. “Anoda” Rodowicz, J. Anoda Rodowicz, “Anoda” Rodowicz, Anoda Rodowicz, Jan Rodowicz, J. Rodowicz, Rodowicz

3.1 Morphological Description of Components

Clearly, most variants of a multi-word unit can be obtained by combinations of inflected forms of their components. Therefore the morphological description of individual words, both common words and proper names, is a necessary prerequisite for our dictionary. In *Toposław* this description is ensured by *Morfeusz SGJP*, a morphological analyser and generator for Polish single words based on the data of the *Grammatical Dictionary of Polish* Saloni *et al.* (2007). For the needs of our dictionary *Morfeusz SGJP* (further called *Morfeusz*) has been extended with 1612 entries: 1005 nouns (not surprisingly last names and geographical names) and

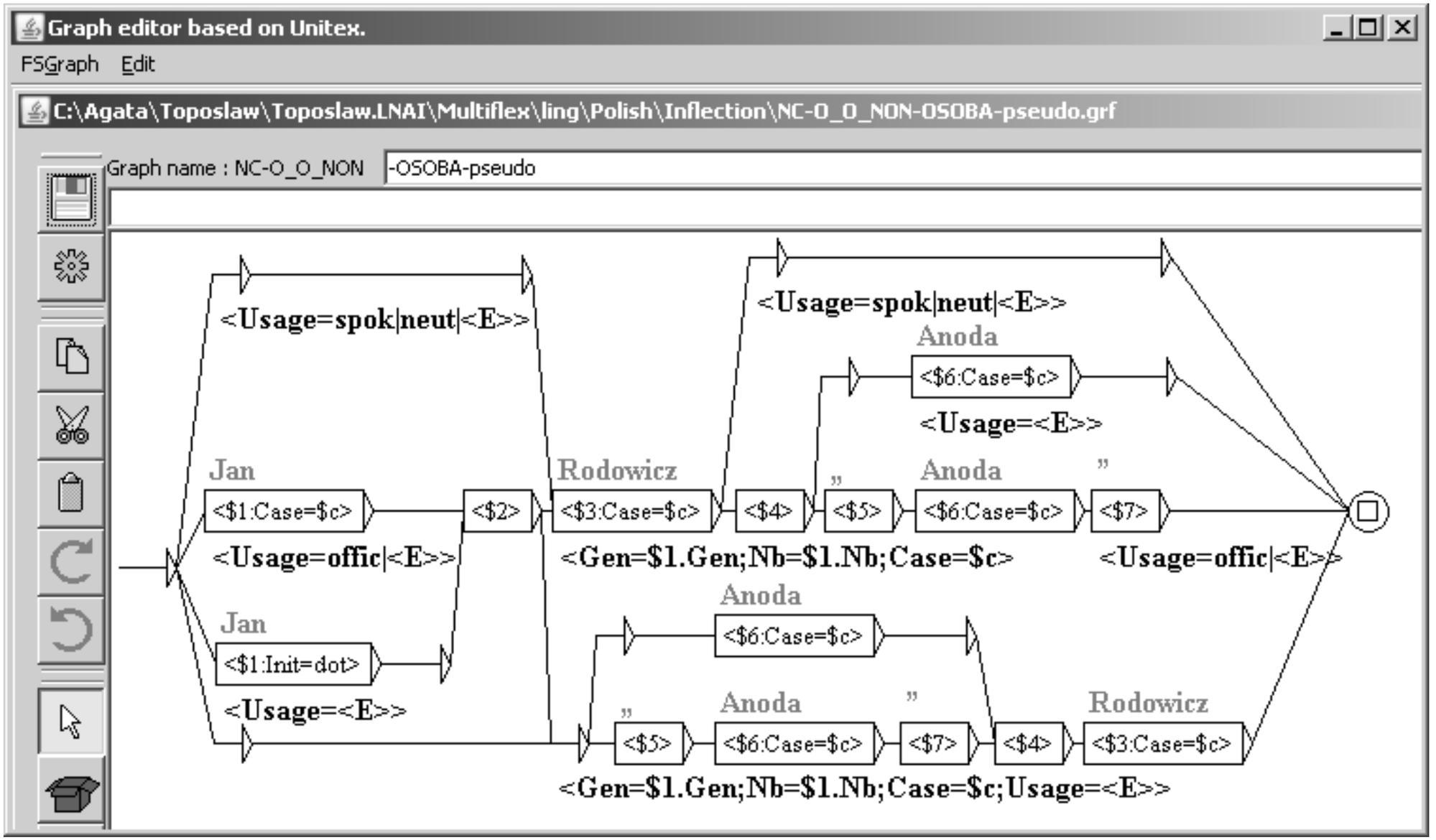


FIGURE 3: Inflection graph for the name *Jan Rodowicz „Anoda"*

feature values, a feature structure can contain alternative values. Here, the alternative operator '—' allows us to reduce the graph's size from 26 to 15 paths. Note that two different values of the same category cannot be selected on the same path. Here for instance the form *Jan Anoda Rodowicz* can be generated with the unmarked value of *Usage* but not with *offic* because the path omitting quotes (i.e. components *$5* and *$7*) has the constraint ⟨*Usage*=⟨*E*⟩⟩.

3.2.3 Initials and Letter Case

Urban proper names frequently take abbreviated forms of their components when appearing in written texts. Any first name can be reduced to its one or two initial letters followed by a dot as in example (1). Similar behavior can be noted in the words 'Street', 'Square', 'Avenue', as well as titles and functions such as 'General'.

Using a dictionary of abbreviations would be most appropriate, unfortunately we are not aware of such a dictionary for Polish. Thus we propose the following partial solution. Whenever an abbreviation is constructed from one to five initial letters, whether or not followed by a dot, the category *Init* (mentioned in Fig. 5) with the corresponding value can be used. For instance in Fig. 3 component *$1* (*Jan*) can be replaced by its initial (*J.*) due to the equation *Init=dot*.

Note that some words are abbreviated differently than by a prefix, as in *płk* for *pułkownik* 'colonel'. Moreover abbreviations or acronyms may be formed from inflected words as *W-wie* for *Warszawie*, or may become independent inflecting lexemes, e.g. *ONZ*, *ONZ-u* for *Organizacja Narodów Zjednoczonych* 'United Nations Organization'. Such cases are currently described with specific dedicated inflectional graphs.

Many urban names contain capitalized common words, as for instance *ulica* ***Długa*** 'Long Str.' or ***Most*** *Syreny* 'Siren Bridge' These words are described in Morfeusz in lower case only. Thus, when such components are morphologically analyzed they obtain lower case lemmas. For instance in Fig. 1 the nickname *Anoda* is assigned a common word lemma *anoda* 'anode'. In order to express this difference in spelling between the lemma and its form we have introduced the *LetterCase* category (see Fig. 6) mentioned in Fig. 5. It indicates how to transform the letter case of the lemma into the form desired in the dictionary. If no transformation is needed the value is *same*.

The *LetterCase* value is most often implicit, i.e. it does not appear in inflection graphs but is automatically deduced from each component of a compound during its morphological analysis. It can however also be explicitly used in graphs if needed.

4 Graph Management

Graphs are created and modified using an enhanced graph editor based on *Unitex* Paumier (2003). It allows for the creation, connection, filling out and deletion of boxes. A graph can be assigned to one or many entries at a time whenever they are simultaneously selected from the list of names.

As the number of graphs grows with the number of names described, managing graphs becomes difficult. Currently we have over 8,900 names with 451

Form	**Lemma**	**LetterCase**
Władysław	*Władysław*	*same*
empik	*EMPiK*	*all_lower*
STUDIO	*studio*	*all_upper*
Długa	*długi*	*first_upper*
Centrum handlowe	*centrum handlowe*	*first_upper*
Centrum Handlowe	*centrum handlowe*	*first_upper_each_word*
1976	*1976*	*no_letter_case*
MarcPol	*Marcpol*	*other*

FIGURE 6: Values of the *LetterCase* category

corresponding graphs. The majority of names use only a few graphs, which are thus easy to remember. For the rest, however, the user needs some support.

4.1 Filtering Graphs

When a lexicographer introduces a new proper name, *Toposław* displays the list of currently defined graphs which have the same number of components as the name in question. This significantly reduces the number of graphs that have to be considered.

Moreover in the process of describing a compound, the user is asked to state for each component whether it inflects or not (see the *Inflects* field in the *Constituents* panel in Fig. 1). This pattern of inflecting components is again used to filter the list of graphs. Namely, a graph matches a name only if:

- for each component marked as inflecting there exists a corresponding box in the graph marked as inflecting (with an equality on some grammatical feature),
- for each component marked as non-inflecting none of the corresponding boxes allow for inflection.

We plan to extend this mechanism by measuring the morphological similarity of components of a name being considered to the components of names already described. This measure will allow us to rank the graphs and suggest the most promising graph for a given name (see Krstev and Vitas (2009) and Krstev *et al.* (2010) for other graph prediction facilities).

One of the novel features introduced in our version of Unitex's graph editor is the labeling of boxes in the graph with constituents of the compound. When the user selects a graph from the list, a preview of this graph is displayed. As shown in Fig. 3 and Fig. 4, all boxes in this preview image are labelled with the constituents of the current name, which allows one to check, whether the components fit into the graph.

4.2 Tracing Paths in a Graph

After a graph is assigned to a name, its inflection is validated by generating and checking all possible forms, as shown in Fig. 7. If an erroneous form is generated, the lexicographer has to deduce, how this form was obtained. For some names,

lexemes, e.g. *ONZ*, *ONZ-u* for *Organizacja Narodów Zjednoczonych* 'United Nations Organization'. Such cases are currently described with specific dedicated inflectional graphs.

Many urban names contain capitalized common words, as for instance *ulica* ***Długa*** 'Long Str.' or ***Most*** *Syreny* 'Siren Bridge' These words are described in Morfeusz in lower case only. Thus, when such components are morphologically analyzed they obtain lower case lemmas. For instance in Fig. 1 the nickname *Anoda* is assigned a common word lemma *anoda* 'anode'. In order to express this difference in spelling between the lemma and its form we have introduced the *LetterCase* category (see Fig. 6) mentioned in Fig. 5. It indicates how to transform the letter case of the lemma into the form desired in the dictionary. If no transformation is needed the value is *same*.

The *LetterCase* value is most often implicit, i.e. it does not appear in inflection graphs but is automatically deduced from each component of a compound during its morphological analysis. It can however also be explicitly used in graphs if needed.

4 Graph Management

Graphs are created and modified using an enhanced graph editor based on *Unitex* Paumier (2003). It allows for the creation, connection, filling out and deletion of boxes. A graph can be assigned to one or many entries at a time whenever they are simultaneously selected from the list of names.

As the number of graphs grows with the number of names described, managing graphs becomes difficult. Currently we have over 8,900 names with 451

corresponding graphs. The majority of names use only a few graphs, which are thus easy to remember. For the rest, however, the user needs some support.

4.1 Filtering Graphs

When a lexicographer introduces a new proper name, *Toposław* displays the list of currently defined graphs which have the same number of components as the name in question. This significantly reduces the number of graphs that have to be considered.

Moreover in the process of describing a compound, the user is asked to state for each component whether it inflects or not (see the *Inflects* field in the *Constituents* panel in Fig. 1). This pattern of inflecting components is again used to filter the list of graphs. Namely, a graph matches a name only if:

- for each component marked as inflecting there exists a corresponding box in the graph marked as inflecting (with an equality on some grammatical feature),
- for each component marked as non-inflecting none of the corresponding boxes allow for inflection.

We plan to extend this mechanism by measuring the morphological similarity of components of a name being considered to the components of names already described. This measure will allow us to rank the graphs and suggest the most promising graph for a given name (see Krstev and Vitas (2009) and Krstev *et al.* (2010) for other graph prediction facilities).

One of the novel features introduced in our version of Unitex's graph editor is the labeling of boxes in the graph with constituents of the compound. When the user selects a graph from the list, a preview of this graph is displayed. As shown in Fig. 3 and Fig. 4, all boxes in this preview image are labelled with the constituents

path. Here for instance the form *Jan Anoda Rodowicz* can be generated with the unmarked value of *Usage* but not with *offic* because the path omitting quotes (i.e. components *$5* and *$7*) has the constraint ⟨*Usage*=⟨*E*⟩⟩.

3.2.3 Initials and Letter Case

Urban proper names frequently take abbreviated forms of their components when appearing in written texts. Any first name can be reduced to its one or two initial letters followed by a dot as in example (1). Similar behavior can be noted in the words 'Street', 'Square', 'Avenue', as well as titles and functions such as 'General'.

Using a dictionary of abbreviations would be most appropriate, unfortunately we are not aware of such a dictionary for Polish. Thus we propose the following partial solution. Whenever an abbreviation is constructed from one to five initial letters, whether or not followed by a dot, the category *Init* (mentioned in Fig. 5) with the corresponding value can be used. For instance in Fig. 3 component *$1* (*Jan*) can be replaced by its initial (*J.*) due to the equation *Init=dot*.

Note that some words are abbreviated differently than by a prefix, as in *płk* for *pułkownik* 'colonel'. Moreover abbreviations or acronyms may be formed from inflected words as *W-wie* for *Warszawie*, or may become independent inflecting lexemes, e.g. *ONZ*, *ONZ-u* for *Organizacja Narodów Zjednoczonych* 'United Nations Organization'. Such cases are currently described with specific dedicated inflectional graphs.

Many urban names contain capitalized common words, as for instance *ulica* ***Długa*** 'Long Str.' or ***Most*** *Syreny* 'Siren Bridge' These words are described in Morfeusz in lower case only. Thus, when such components are morphologically analyzed they obtain lower case lemmas. For instance in Fig. 1 the nickname *Anoda* is assigned a common word lemma *anoda* 'anode'. In order to express this difference in spelling between the lemma and its form we have introduced the *LetterCase* category (see Fig. 6) mentioned in Fig. 5. It indicates how to transform the letter

ELSEVIER

Artificial Intelligence 78 (1995) 87–119

Artificial Intelligence

A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry

Zhengyou Zhang *, Rachid Deriche, Olivier Faugeras, Quang-Tuan Luong

INRIA Sophia-Antipolis, 2004 route des Lucioles, B.P. 93, F-06902 Sophia-Antipolis Cedex, France

Received June 1994; revised December 1994

Abstract

This paper proposes a robust approach to image matching by exploiting the only available geometric constraint, namely, the epipolar constraint. The images are uncalibrated, namely the motion between them and the camera parameters are not known. Thus, the images can be taken by different cameras or a single camera at different time instants. If we make an exhaustive search for the epipolar geometry, the complexity is prohibitively high. The idea underlying our approach is to use classical techniques (correlation and relaxation methods in our particular implementation) to find an initial set of matches, and then use a robust technique—the Least Median of Squares (LMedS)—to discard false matches in this set. The epipolar geometry can then be accurately estimated using a meaningful image criterion. More matches are eventually found, as in stereo matching, by using the recovered epipolar geometry. A large number of experiments have been carried out, and very good results have been obtained.

Regarding the relaxation technique, we define a new measure of matching support, which allows a higher tolerance to deformation with respect to rigid transformations in the image plane and a smaller contribution for distant matches than for nearby ones. A new strategy for updating matches is developed, which only selects those matches having both high matching support and low matching ambiguity. The update strategy is different from the classical "winner-take-all", which is easily stuck at a local minimum, and also from "loser-take-nothing", which is usually very slow. The proposed algorithm has been widely tested and works remarkably well in a scene with many repetitive patterns.

Keywords: Robust matching; Epipolar geometry; Fundamental matrix; Least Median of Squares (LMedS); Relaxation; Correlation

* Corresponding author. E-mail: zzhang@@sophia.inria.fr.

SSDI 0004-3702(95)00022-4

which computes an initial estimate of the epipolar geometry over all matches, and sees how the estimate changes if a match is deleted. The match whose removal maximally reduces the residual is identified to be an *outlier* and is rejected. The procedure is then repeated with the reduced set of matches until all outliers have been removed. These two approaches (M-estimators and regression diagnostics) work well when the percentage of outliers is small and more importantly when their derivations from the valid matches are not too large, as in the above two works. In the case described in this paper, two images can be quite different. There may be a large percentage of false matches (usually around 20%, sometimes 40%) using heuristic matching techniques such as correlation, and a false match may be completely different from the valid matches. The robust technique described in this paper deals with these issues and can theoretically detect outliers when they make up as much as 50% of whole data.

2. Notation

A camera is described by the widely used pinhole model. The coordinates of a 3D point $M = [x, y, z]^T$ in a world coordinate system and its retinal image coordinates $\boldsymbol{m} = [u, v]^T$ are related by

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbb{P} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix},$$

where s is an arbitrary scale, and $\mathbb{P}$ is a 3×4 matrix, called the perspective projection matrix. Denoting the homogeneous coordinates of a vector $\boldsymbol{x} = [x, y, \ldots]^T$ by $\tilde{\boldsymbol{x}}$, i.e., $\tilde{\boldsymbol{x}} = [x, y, \ldots, 1]^T$, we have $s\tilde{\boldsymbol{m}} = \mathbb{P}\tilde{M}$.

The matrix $\mathbb{P}$ can be decomposed as

$$\mathbb{P} = \boldsymbol{A}[\boldsymbol{R}\ \boldsymbol{t}],$$

where $\boldsymbol{A}$ is a 3×3 matrix, mapping the normalized image coordinates to the retinal image coordinates, and $(\boldsymbol{R}, \boldsymbol{t})$ is the 3D displacement (rotation and translation) from the world coordinate system to the camera coordinate system. The most general matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \begin{bmatrix} -fk_u & fk_u \cot\theta & u_0 \\ 0 & -fk_v/\sin\theta & v_0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where

- f is the focal length of the camera,
- k_u and k_v are the horizontal and vertical scale factors, whose inverses characterize the size of the pixel in the world coordinate unit,
- u_0 and v_0 are the coordinates of the principal point of the camera, i.e., the intersection between the optical axis and the image plane, and

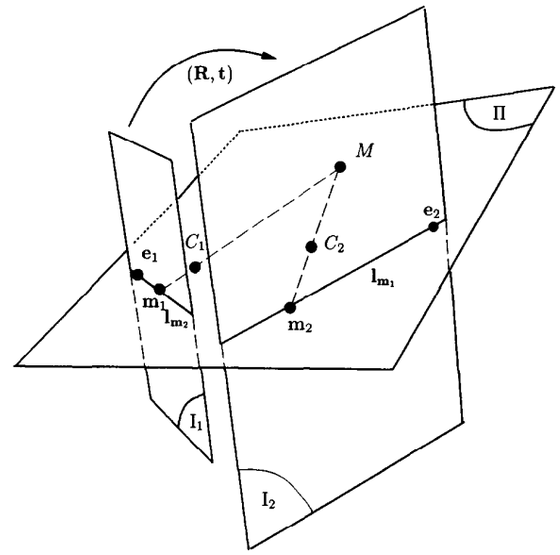


Fig. 1. The epipolar geometry.

- θ is the angle between the retinal axes. (This parameter is introduced to account for the fact that the pixel grid may not be exactly orthogonal. In practice, however, it is very close to $\pi/2$.)

As is clear, we cannot separate f from k_u and k_v. We thus have five intrinsic parameters for each camera: $\alpha_u = -fk_u$, $\alpha_v = -fk_v$, u_0, v_0 and θ.

The first and second images are respectively denoted by I_1 and I_2. A point $\boldsymbol{m}$ in the image plane I_i is noted as $\boldsymbol{m}_i$. The second subscript, if any, will indicate the index of the point in consideration.

3. Epipolar geometry

Considering the case of two cameras as shown in Fig. 1.

Let C_1 and C_2 be the optical centers of the first and second cameras, respectively. Given a point $\boldsymbol{m}_1$ in the first image, its corresponding point in the second image is constrained to lie on a line called the *epipolar line* of $\boldsymbol{m}_1$, denoted by $\boldsymbol{l}_{\boldsymbol{m}_1}$. The line $\boldsymbol{l}_{\boldsymbol{m}_1}$ is the intersection of the plane Π, defined by $\boldsymbol{m}_1$, C_1 and C_2 (known as the *epipolar plane*), with the second image plane I_2. This is because image point $\boldsymbol{m}_1$ may correspond to an arbitrary point on the semi-line C_1M (M may be at infinity) and that the projection of C_1M on I_2 is the line $\boldsymbol{l}_{\boldsymbol{m}_1}$. Furthermore, one observes that all epipolar lines of the points in the first image pass through a common point $\boldsymbol{e}_2$, which is called the *epipole*. $\boldsymbol{e}_2$ is the intersection of the line C_1C_2 with the image plane I_2. This can be easily understood as follows. For each point $\boldsymbol{m}_{1k}$ in the first image I_1, its epipolar line $\boldsymbol{l}_{\boldsymbol{m}_{1k}}$ in

Journal of Machine Learning Research 12 (2011) 491-523 Submitted 6/10; Published 2/11

Learning Multi-modal Similarity

Brian McFee BMCFEE@CS.UCSD.EDU
Department of Computer Science and Engineering
University of California
San Diego, CA 92093-0404, USA

Gert Lanckriet GERT@ECE.UCSD.EDU
Department of Electrical and Computer Engineering
University of California
San Diego, CA 92093-0407, USA

Editor: Tony Jebara

Abstract

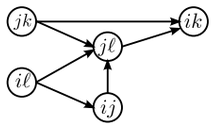
In many applications involving multi-media data, the definition of similarity between items is integral to several key tasks, including nearest-neighbor retrieval, classification, and recommendation. Data in such regimes typically exhibits multiple modalities, such as acoustic and visual content of video. Integrating such heterogeneous data to form a holistic similarity space is therefore a key challenge to be overcome in many real-world applications.

We present a novel multiple kernel learning technique for integrating heterogeneous data into a single, unified similarity space. Our algorithm learns an optimal ensemble of kernel transformations which conform to measurements of human perceptual similarity, as expressed by relative comparisons. To cope with the ubiquitous problems of subjectivity and inconsistency in multi-media similarity, we develop graph-based techniques to filter similarity measurements, resulting in a simplified and robust training procedure.

Keywords: multiple kernel learning, metric learning, similarity

1. Introduction

In applications such as content-based recommendation systems, the definition of a proper similarity measure between items is crucial to many tasks, including nearest-neighbor retrieval and classification. In some cases, a natural notion of similarity may emerge from domain knowledge, for example, cosine similarity for bag-of-words models of text. However, in more complex, multi-media domains, there is often no obvious choice of similarity measure. Rather, viewing different aspects of the data may lead to several different, and apparently equally valid notions of similarity. For example, if the corpus consists of musical data, each song or artist may be represented simultaneously by acoustic features (such as rhythm and timbre), semantic features (tags, lyrics), or social features (collaborative filtering, artist reviews and biographies, etc). Although domain knowledge may be incorporated to endow each representation with an intrinsic geometry—and, therefore, a sense of similarity—the different notions of similarity may not be mutually consistent. In such cases, there is generally no obvious way to combine representations to form a unified similarity space which optimally integrates heterogeneous data.



$$\mathcal{C} = \left\{ \begin{array}{ll} (j,k,j,\ell), & (j,k,i,k), \\ (j,\ell,i,k), & (i,\ell,j,\ell), \\ (i,\ell,i,j), & (i,j,j,\ell) \end{array} \right\}$$

Figure 2: The graph representation (left) of a set of relative comparisons (right).

1. If $\mathcal{C}$ contains cycles, then there exists no embedding which can satisfy $\mathcal{C}$.
2. If $\mathcal{C}$ is acyclic, any embedding that satisfies the transitive reduction $\mathcal{C}^{\min}$ also satisfies $\mathcal{C}$.

The first fact implies that no algorithm can produce an embedding which satisfies all measurements if the graph is cyclic. In fact, the converse of this statement is also true: if $\mathcal{C}$ is acyclic, then an embedding exists in which all similarity measurements are preserved (see Appendix A). If $\mathcal{C}$ is cyclic, however, by analyzing the graph, it is possible to identify an "unlearnable" subset of $\mathcal{C}$ which must be violated by any embedding.

Similarly, the second fact exploits the transitive nature of distance comparisons. In the example depicted in Figure 2, any g that satisfies (j,k,j,ℓ) and (j,ℓ,i,k) must also satisfy (j,k,i,k). In effect, the constraint (j,k,i,k) is redundant, and may also be safely omitted from $\mathcal{C}$.

These two observations allude to two desirable properties in $\mathcal{C}$ for embedding methods: *transitivity* and *anti-symmetry*. Together with irreflexivity, these fit the defining characteristics of a *partial order*. Due to subjectivity and inter-labeler disagreement, however, most collections of relative comparisons will not define a partial order. Some graph processing, presented next, based on an approximate maximum acyclic subgraph algorithm, can reduce them to a partial order.

2.2 Graph Simplification

Because a set of similarity measurements $\mathcal{C}$ containing cycles cannot be embedded in any Euclidean space, $\mathcal{C}$ is inherently inconsistent. Cycles in $\mathcal{C}$ therefore constitute a form of *label noise*. As noted by Angelova (2004), label noise can have adverse effects on both model complexity and generalization. This problem can be mitigated by detecting and pruning noisy (confusing) examples, and training on a reduced, but certifiably "clean" set (Angelova et al., 2005; Vezhnevets and Barinova, 2007).

Unlike most settings, where the noise process affects each label independently—for example, random classification noise (Angluin and Laird, 1988)—the graphical structure of interrelated relative comparisons can be exploited to detect and prune inconsistent measurements. By eliminating similarity measurements which cannot be realized by any embedding, the optimization procedure can be carried out more efficiently and reliably on a reduced constraint set.

Ideally, when eliminating edges from the graph, we would like to retain as much information as possible. Unfortunately, this is equivalent to the *maximum acyclic subgraph* problem, which is NP-Complete (Garey and Johnson, 1979). A $1/2$-approximate solution can be achieved by a simple greedy algorithm (Algorithm 1) (Berger and Shor, 1990).

Once a consistent subset of similarity measurements has been produced, it can be simplified further by pruning redundancies. In the graph view of similarity measurements, redundancies can be easily removed by computing the transitive reduction of the graph (Aho et al., 1972).

Algorithm 1 Approximate maximum acyclic subgraph (Aho et al., 1972)

Input: Directed graph $G = (V, E)$
Output: Acyclic graph G'
$E' \leftarrow \emptyset$
for each $(u,v) \in E$ in random order **do**
  if $E' \cup \{(u,v)\}$ is acyclic **then**
    $E' \leftarrow E' \cup \{(u,v)\}$
  end if
end for
$G' \leftarrow (V, E')$

By filtering the constraint set for consistency, we ensure that embedding algorithms are not learning from spurious information. Additionally, pruning the constraint set by transitive reduction focuses embedding algorithms on the most important core set of constraints while reducing overhead due to redundant information.

3. Partial Order Embedding

Now that we have developed a language for expressing similarity between items, we are ready to formulate the embedding problem. In this section, we develop an algorithm that learns a representation of data consistent with a collection of relative similarity measurements, and allows to map unseen data into the learned similarity space after learning. In order to accomplish this, we will assume a feature representation for $\mathcal{X}$. By parameterizing the embedding function g in terms of the feature representation, we will be able to apply g to any point in the feature space, thereby generalizing to data outside of the training set.

3.1 Linear Projection

To start, we assume that the data originally lies in some Euclidean space, that is, $\mathcal{X} \subset \mathbb{R}^D$. There are of course many ways to define an embedding function $g : \mathbb{R}^D \to \mathbb{R}^d$. Here, we will restrict attention to embeddings parameterized by a linear projection matrix M, so that for a vector $x \in \mathbb{R}^D$,

$$g(x) \doteq Mx.$$

Collecting the vector representations of the training set as columns of a matrix $X \in \mathbb{R}^{D \times n}$, the inner product matrix of the embedded points can be characterized as

$$A = X^{\mathsf{T}} M^{\mathsf{T}} M X.$$

Now, for a relative comparison (i,j,k,ℓ), we can express the distance constraint (1) between embedded points as follows:

$$(X_i - X_j)^{\mathsf{T}} M^{\mathsf{T}} M (X_i - X_j) + 1 \leq (X_k - X_\ell)^{\mathsf{T}} M^{\mathsf{T}} M (X_k - X_\ell).$$

These inequalities can then be used to form the constraint set of an optimization problem to solve for M. Because, in general, $\mathcal{C}$ may not be satisfiable by a linear projection of $\mathcal{X}$, we soften the

Finite-State Chart Constraints for Reduced Complexity Context-Free Parsing Pipelines

Brian Roark*
Oregon Health & Science University

Kristy Hollingshead**
University of Maryland

Nathan Bodenstab*
Oregon Health & Science University

We present methods for reducing the worst-case and typical-case complexity of a context-free parsing pipeline via hard constraints derived from finite-state pre-processing. We perform $O(n)$ predictions to determine if each word in the input sentence may begin or end a multi-word constituent in chart cells spanning two or more words, or allow single-word constituents in chart cells spanning the word itself. These pre-processing constraints prune the search space for any chart-based parsing algorithm and significantly decrease decoding time. In many cases cell population is reduced to zero, which we term chart cell "closing." We present methods for closing a sufficient number of chart cells to ensure provably quadratic or even linear worst-case complexity of context-free inference. In addition, we apply high precision constraints to achieve large typical-case speedups and combine both high precision and worst-case bound constraints to achieve superior performance on both short and long strings. These bounds on processing are achieved without reducing the parsing accuracy, and in some cases accuracy improves. We demonstrate that our method generalizes across multiple grammars and is complementary to other pruning techniques by presenting empirical results for both exact and approximate inference using the exhaustive CKY algorithm, the Charniak parser, and the Berkeley parser. We also report results parsing Chinese, where we achieve the best reported results for an individual model on the commonly reported data set.

1. Introduction

Although there have been great advances in the statistical modeling of hierarchical syntactic structure over the past 15 years, exact inference with such models remains very costly and most rich syntactic modeling approaches resort to heavy pruning, pipelining,

* Center for Spoken Language Understanding, Oregon Health & Science University, Beaverton, Oregon, 97006 USA. E-mails: roarkbr@gmail.com, bodenstab@gmail.com.
** Some of the work in this paper was done while Kristy Hollingshead was at OHSU. She is currently at the University of Maryland Institute for Advanced Computer Studies, College Park, Maryland, 20740 USA. E-mail: hollingk@gmail.com.

Submission received: 9 August 2011; revised submission received: 30 November 2011; accepted for publication: 4 January 2012.

Table 2
Tagger features for $\overline{B}$, $\overline{E}$, and $\overline{U}$.

	LEX	ORTHO	POS
τ_i	τ_i, w_i	$\tau_i, w_i[0]$	τ_i, φ_i
τ_{i-1}, τ_i	τ_i, w_{i-1}	$\tau_i, w_i[0..1]$	τ_i, φ_{i-1}
τ_{i-2}, τ_i	τ_i, w_{i+1}	$\tau_i, w_i[0..2]$	$\tau_i, \varphi_{i-1}, \varphi_i$
$\tau_{i-2}, \tau_{i-1}, \tau_i$	τ_i, w_{i-2}	$\tau_i, w_i[0..3]$	τ_i, φ_{i+1}
	τ_i, w_{i+2}	$\tau_i, w_i[n]$	$\tau_i, \varphi_i, \varphi_{i+1}$
	τ_i, w_{i-1}, w_i	$\tau_i, w_i[n\text{-}1..n]$	$\tau_i, \varphi_{i-1}, \varphi_i, \varphi_{i+1}$
	τ_i, w_i, w_{i+1}	$\tau_i, w_i[n\text{-}2..n]$	τ_i, φ_{i-2}
		$\tau_i, w_i[n\text{-}3..n]$	$\tau_i, \varphi_{i-2}, \varphi_{i-1}$
		$\tau_i, w_i \subseteq$ Digit	$\tau_i, \varphi_{i-2}, \varphi_{i-1}, \varphi_i$
		$\tau_i, w_i \subseteq$ UpperCase	τ_i, φ_{i+2}
		$\tau_i, w_i \subseteq$ Hyphen	$\tau_i, \varphi_{i+1}, \varphi_{i+2}$
			$\tau_i, \varphi_i, \varphi_{i+1}, \varphi_{i+2}$

All lexical (LEX), orthographic (ORTHO), and part-of-speech (POS) features are duplicated to also occur with τ_{i-1}; e.g., $\{\tau_{i-1}, \tau_i, w_i\}$ as a LEX feature.

the preceding words. The *n*-gram features are represented by the words within a three-word window of the current word. The tag features are represented as unigram, bigram, and trigram tags (i.e., constituent tags from the current and two previous words). These features are based on the feature set implemented by Sha and Pereira (2003) for NP chunking. Additional orthographical features are used for unknown and rare words (words that occur fewer than five times in the training data), such as the prefixes and suffixes of the word (up to the first and last four characters of the word), and the presence of a hyphen, a digit, or a capitalized letter, following the features implemented by Ratnaparkhi (1999). Note that the orthographic feature templates, including the prefix (e.g., $w_i[0..1]$) and suffix (e.g., $w_i[n\text{-}2..n]$) templates, are only activated for unknown and rare words. When applying our tagging model to Chinese data, all feature functions were left in the model as-is, and not tailored to the specifics of the language.

We ran various tagging experiments on the development set and report accuracy results in Table 3 for all three predictions tasks, using Viterbi decoding. We trained

Table 3
Tagging accuracy on the respective development sets (WSJ Section 24 for English and Penn Chinese Treebank articles 301–325 for Chinese) for binary classes $\overline{B}$, $\overline{E}$, and $\overline{U}$, for various Markov orders.

Tagging Task	**Markov order**		
	0	**1**	**2**
English			
$\overline{B}$ (no multi-word constituent begin)	96.7	96.9	96.9
$\overline{E}$ (no multi-word constituent end)	97.3	97.3	97.3
$\overline{U}$ (no span-1 unary constituent)	98.3	98.3	98.3
Chinese			
$\overline{B}$ (no multi-word constituent begin)	94.8	95.4	95.2
$\overline{E}$ (no multi-word constituent end)	96.2	96.4	96.6
$\overline{U}$ (no span-1 unary constituent)	95.9	96.2	96.3

models with Markov order-0 (constituent tags predicted for each word independently), order-1 (features with constituent tag pairs), and order-2 (features with constituent tag triples). In general, tagging accuracy for English is higher than for Chinese, especially for the $\overline{U}$ and $\overline{B}$ tasks. Given the consistent improvement from Markov order-0 to Markov order-1 (particularly on the Chinese data), and for the sake of consistency, we have chosen to perform Markov order-1 prediction for all results in the remainder of this article.

6. Experimental Set-up

In the sections that follow, we present empirical trials to examine the behavior of chart constraints under a variety of conditions. First, we detail the data, evaluation, and parsers used in these experiments.

6.1 Data Sets and Evaluation

For English, all stochastic grammars are induced from the Penn WSJ Treebank (Marcus, Marcinkiewicz, and Santorini 1993). Sections 2–21 of the treebank are used as training, Section 00 as held-out (for determining stopping criteria during training and some parameter tuning), Section 24 as development, and Section 23 as test set. For Chinese, we use the Penn Chinese Treebank (Xue et al. 2005). Articles 1–270 and 400–1151 are used for training, articles 301–325 for both held-out and development, and articles 271–300 for testing. Supervised class labels are extracted from the non-binarized treebank trees for B, E, and U (as well as their complements).

All results report F-measure labeled bracketing accuracy (harmonic mean of labeled precision and labeled recall) for all sentences in the data set (Black et al. 1991), and timing is reported using an Intel 3.00GHz processor with 6MB of cache and 16GB of memory. Timing results include both the pre-processing time to tag the chart constraints as well as the subsequent context-free inference, but tagging time is relatively negligible as it takes less than three seconds to tag the entire development corpus.

6.2 Tagging Methods and Closing Chart Cells

We have three separate tagging tasks, each with two possible tags for every word w_i in the input string: (1) B or $\overline{B}$; (2) E or $\overline{E}$; and (3) U or $\overline{U}$. Our taggers are as described in Section 5.

Within a pipeline system that leverages hard constraints, one may want to choose a tagger operating point that favors precision of constraints over recall to avoid over-constraining the downstream parser. We have two methods for trading recall for precision that will be detailed later in this section, both relying on calculating the cumulative score S_i for each of the binary tags at each word position w_i. That is, (using B as the example tag):

$$S_i(B \mid w_1 \dots w_n) = \log \sum_{\tau_1 \dots \tau_n} \delta(\tau_i, B)\, e^{\Phi(w_1 \dots w_n, \tau_1 \dots \tau_n) \cdot \mathbf{w}} \quad (1)$$

where $\sum$ sums over all possible tag sequences for sentence $w_1 \dots w_n$; $\delta(\tau_i, B) = 1$ if $\tau_i = B$ and 0 otherwise; $\Phi(w_1 \dots w_n, \tau_1 \dots \tau_n)$ maps the word string and particular tag string to a d-dimensional (global) feature vector; and $\mathbf{w}$ is the d-dimensional parameter vector

Toposław II

Toposław — a lexicographic framework for multi-word units

Małgorzata Marciniak[1], *Agata Savary*[2,1]
Piotr Sikora[1] *and Marcin Woliński*[1]

ABSTRACT

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morphosyntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 INTRODUCTION

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

We are interested in a lexical description of multi-word names in a particular application domain: the urban transportation system in Warsaw. In Marciniak *et al.*(2009) we present a project of creating a dictionary of Polish toponyms relevant to Warsaw transportation. The project includes the development of a lexicographic framework, *Toposław*, that allows us the creation of the dictionary in an efficient and quality-ensured manner. *Toposław*c ooperates with: (i) *Morfeusz SGJP* (a new version of *Morfeusz* Woliński (2006)), a morphological analyzer and generator for Polish single words, (ii) *Multiflex* Savary (2005), a cross-language, mor-

1 Institute of Computer Science, Polish Academy of Sciences, J.K. Ordona 21, 01-237 Warszawa, Poland

2 Université François Rabelais Tours, 3, pl. Jean-Jaurès, 41029 Blois, France

Toposław — a lexicographic framework for multi-word units

Małgorzata Marciniak[1], Agata Savary[2,1]
Piotr Sikora[1] and Marcin Woliński[1]

ABSTRACT

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morphosyntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 INTRODUCTION

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

We are interested in a lexical description of multi-word names in a particular application domain: the urban transportation system in Warsaw. In Marciniak *et al.*(2009) we present a project of creating a dictionary of Polish toponyms relevant to Warsaw transportation. The project includes the development of a lexicographic framework, *Toposław*, that allows us the creation of the dictionary in an efficient and quality-ensured manner. *Toposław*c ooperates with: (i) *Morfeusz SGJP* (a new version of *Morfeusz* Woliński (2006)), a morphological analyzer and generator for Polish single words, (ii) *Multiflex* Savary (2005), a cross-language, mor-

1 Institute of Computer Science, Polish Academy of Sciences, J.K. Ordona 21, 01-237 Warszawa, Poland

2 Université François Rabelais Tours, 3, pl. Jean-Jaurès, 41029 Blois, France

Toposław – a lexicographic framework for multi-word units

Małgorzata Marciniak[1], Agata Savary[2,1]
Piotr Sikora[1] and Marcin Woliński[1]

Abstract

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morphosyntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 Introduction

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

We are interested in a lexical description of multi-word names in a particular application domain: the urban transportation system in Warsaw. In Marciniak *et al.*(2009) we present a project of creating a dictionary of Polish toponyms relevant to Warsaw transportation. The project includes the development of a lexicographic framework, *Toposław*, that allows us the creation of the dictionary in an efficient and quality-ensured manner. *Toposław*c ooperates with: (i) *Morfeusz SGJP* (a new version of *Morfeusz* Woliński (2006)), a morphological analyzer and generator for Polish single words, (ii) *Multiflex* Savary (2005), a cross-language, mor-

1 Institute of Computer Science, Polish Academy of Sciences, J.K. Ordona 21, 01-237 Warszawa, Poland

2 Université François Rabelais Tours, 3, pl. Jean-Jaurès, 41029 Blois, France

Toposław — a lexicographic framework for multi-word units

Małgorzata Marciniak[1], *Agata Savary*[2,1]
Piotr Sikora[1] *and Marcin Woliński*[1]

ABSTRACT

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morphosyntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 INTRODUCTION

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction

1 Institute of Computer Science, Polish Academy of Sciences, J.K.

Toposław – a lexicographic framework for multi-word units

Małgorzata Marciniak[1], *Agata Savary*[2,1]
Piotr Sikora[1] *and Marcin Woliński*[1]

Abstract

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morphosyntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords: electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 Introduction

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names

1 Institute of Computer Science, Polish Academy of Sciences, J.K.

JLM vol. 0 issue 1

Design spec

general layout
main page objects
typography
title page

Technology

prototyped in Adobe InDesign
implemented in XeLaTeX
PDF as output format
open-source fonts

Part 2: Example layouts
(Draft / preliminary, not 100% conforming to the spec)

Toposław — a lexicographic framework for multi-word units

Małgorzata Marciniak[1], Agata Savary[2]
Piotr Sikora[1] and Marcin Woliński[1]

1 Institute of Computer Science, Polish Academy of Sciences, J.K. Ordona 21, 01-237 Warszawa, Poland
2 Université François Rabelais Tours, 3, pl. Jean-Jaurès, 41029 Blois, France

ABSTRACT

The paper presents a tool for the creation of an electronic dictionary of multiword proper names. *Toposław* uses graphs for the representation of inflectional and pragmatic variants of names. It cooperates with *Morfeusz*, a morphological analyser and generator for Polish words, and *Multiflex*, a cross-language morpho-syntactic generator of multi-word units. Our goal was to create a user-friendly tool that makes a lexicographic work easy and efficient. In the paper we describe facilities for graph creation, management and debugging. The presented tool was applied to create a dictionary of Warsaw urban proper names.

Keywords:
electronic dictionary, lexicographic framework, graphs, urban proper names, Polish

1 INTRODUCTION

Proper names and other named entities are of crucial quantitative and qualitative importance for natural language processing (NLP) due to their frequent occurrence in texts and their rich semantic content. Despite continual efforts in the NLP community aiming at named entity extraction and modelling, the formal linguistic description of proper names remains a challenge.

We are interested in a lexical description of multi-word names in a particular application domain: the urban transportation system in Warsaw. In Marciniak *et al.* (2009) we present a project of creating a dictionary of Polish toponyms relevant

ences, J.K. Ordona 21, 01-237 Warszav

2 Université François Rabelais Tours, 3, pl. Jean-Ja

ABSTRACT

The paper presents a tool for the creation
tionary of multiword proper names. *Toposła*
representation of inflectional and pragmati
It cooperates with *Morfeusz*, a morphologic
erator for Polish words, and *Multiflex*, a cro
syntactic generator of multi-word units. Ou

Sizes and units

Grid units

1 sp = 1 in/72.27/65536

1 mm = 118407168/635 sp

x-axis (width) grid unit

xgu = 895044 sp ≈ 13.606 pt

y-axis (height) grid unit

ygu = 895044 sp ≈ 13.606 pt

```
\newdimen\xgu
\newdimen\ygu
\xgu=895044sp
\ygu=895044sp
```

Sizes and units

PDF media size

170 mm × 240 mm (restricted B5)

Design size

35 xgu × 50 ygu

(∼167.9995 mm × 239.9993 mm)

y dir: top empty area (3 ygu)

y dir: running headers area (1 ygu)

y dir: headers separator (1 ygu)

y dir: main publication flow (38 ygu)

y dir: footers separator (2 ygu)

y dir: running footers area (1 ygu)

y dir: bottom empty area (4 ygu)

{y-direction measurements for all pages}

x dir: inside margin (4 xgu)

Article Title
(23 xgu x 4 ygu centred)

one or two columns for authors with affiliations

one or two columns for authors with affiliations

Main text starts here.
For Article Title page, the main text frame should be:
x dir: 23 + 5 standard width frame (23 xgu)
y dir: offset it by 12 ygu from start of main flow,
i.e. height = 38 – 12 = 26 ygu

(Main text)

x dir: 23 + 5 standard width sidebar (5 xgu)

(Side annotations)

x dir: outside margin (2 xgu)

{title page layout; note: no headers or footers!}

y dir: top empty area (3 ygu)

y dir: running headers area (1 ygu)

y dir: headers separator (1 ygu)

y dir: main publication flow (38 ygu)

y dir: footers separator (2 ygu)

y dir: running footers area (1 ygu)

y dir: bottom empty area (4 ygu)

{y-direction measurements for all pages}

x dir: inside margin (4 xgu)

x dir: outside margin (2 xgu)

Article Title

(23 xgu x 4 ygu centred)

one or two columns for authors with affiliations

one or two columns for authors with affiliations

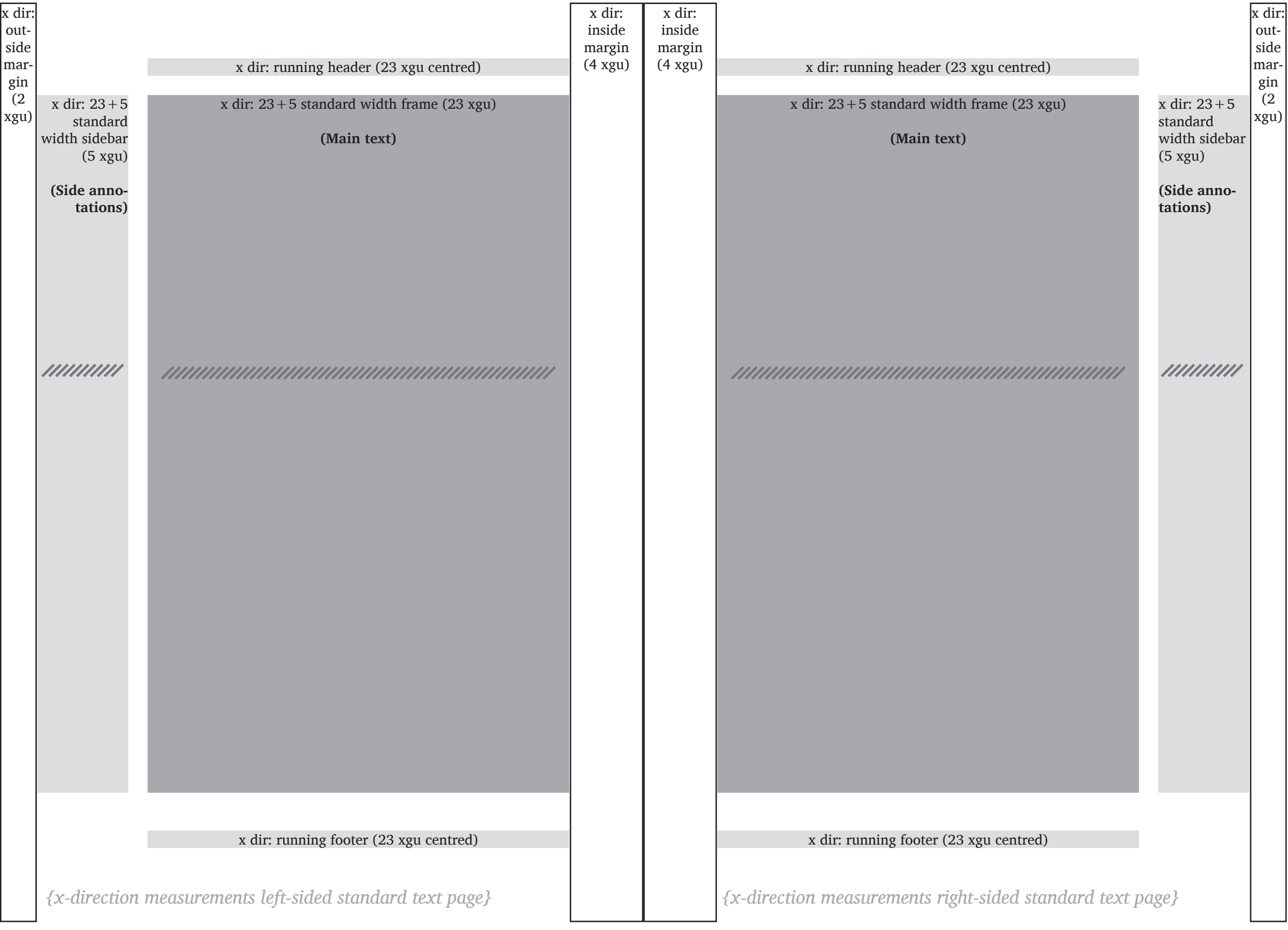
Main text starts here.
For Article Title page, the main text frame should be:
x dir: 23 + 5 standard width frame (23 xgu)
y dir: offset it by 12 ygu from start of main flow,
i.e. height = 38 – 12 = 26 ygu

(Main text)

x dir: 23 + 5 standard width sidebar (5 xgu)

(Side annotations)

{title page layout; note: no headers or footers!}



{x-direction measurements left-sided standard text page}

{x-direction measurements right-sided standard text page}

x dir: outside margin (2 xgu)

x dir: inside margin (4 xgu)

x dir: 23 + standard width sidebar (5 xgu)

(Side annotations)

(Main text)

```
\usepackage[papersize={35\xgu,50\ygu},
    top=3\ygu,includehead,headheight=1\ygu,
    headsep=1\ygu,
    lines=38,
    footskip=3\ygu,
    inner=4\xgu,
    includemp,marginparsep=1\xgu,
    marginparwidth=5\xgu,
    textwidth=23\xgu]
  {geometry}
```

Illustrations

“When images or tables are used, one of the flexible horizontal layouts needs to be applied, depending on the amount of text for the caption and on the proportions of the image. If the image or table is very large or wide, it should occupy the standard width or even the full width. If a caption or annotation is too long to fit into the sidebar, it needs to be placed be underneath the table or image, then occupying the standard width (never the full width!)”

x dir: outside margin (2 xgu)

x dir: inside margin (4 xgu)

x dir: running header (23 xgu centred)

x dir: full width frame (29 xgu)
(Very Large Table or Illustration)

x dir: 23 + 5 standard width sidebar (5 xgu)
(Tiny Caption)

x dir: 23 + 5 standard width frame (23 xgu)
(Large Table or Illustration)

x dir: 20 + 6 sidebar (6 xgu)

(Short Caption)

x dir: 20 + 6 frame (20 xgu)
(Largeish Table or Illustration)

x dir: 17 + 9 sidebar (9 xgu)

(Medium Caption)

x dir: 17 + 9 frame (17 xgu)
(Medium Table or Illustration)

x dir: 14 + 14 full width two-column frame (14 xgu)

(Long Caption)

x dir: 14 + 14 full width two-column frame (14 xgu)
(Small Illustration)

x dir: 11 + 17 full width two-column frame (17 xgu)

(Huge Caption)

x dir: 11 + 17 full width two-column frame (11 xgu)
(Tiny Illustration)

x dir: running footer (23 xgu centred)

{x-direction measurements illustrations, left-sided page}

x dir: inside margin (4 xgu)

x dir: outside margin (2 xgu)

x dir: running header (23 xgu centred)

x dir: full width frame (29 xgu)
(Very Large Table or Illustration)

x dir: 23 + 5 standard width frame (23 xgu)
(Large Table or Illustration)

x dir: 23 + 5 standard width sidebar (5 xgu)
(Tiny Caption)

x dir: 20 + 6 frame (20 xgu)
(Largeish Table or Illustration)

x dir: 20 + 6 sidebar (6 xgu)

(Short Caption)

x dir: 17 + 9 frame (17 xgu)
(Medium Table or Illustration)

x dir: 17 + 9 sidebar (9 xgu)

(Medium Caption)

x dir: 14 + 14 full width two-column frame (14 xgu)
(Small Illustration)

x dir: 14 + 14 full width two-column frame (14 xgu)

(Long Caption)

x dir: 11 + 17 full width two-column frame (11 xgu)
(Tiny Illustration)

x dir: 11 + 17 full width two-column frame (17 xgu)

(Huge Caption)

x dir: running footer (23 xgu centred)

{x-direction measurements illustrations, right-sided page}

x dir: outside margin (2 xgu)

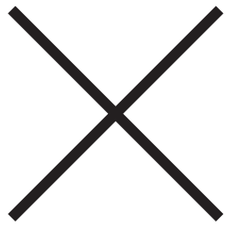
x dir: inside margin (4 xgu)

x dir: running header (23 xgu centred)

x dir: full width frame (29 xgu)
(Very Large Table or Illustration)

x dir: 23 + 5 standard width sidebar (5 xgu)
(Tiny Caption)

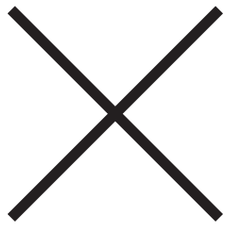
x dir: 23 + 5 standard width frame (23 xgu)
(Large Table or Illustration)



x dir: 20 + 6 sidebar (6 xgu)

(Short Caption)

x dir: 20 + 6 frame (20 xgu)
(Largeish Table or Illustration)



x dir: 17 + 9 sidebar (9 xgu)

x dir: 17 + 9 frame (17 xgu)

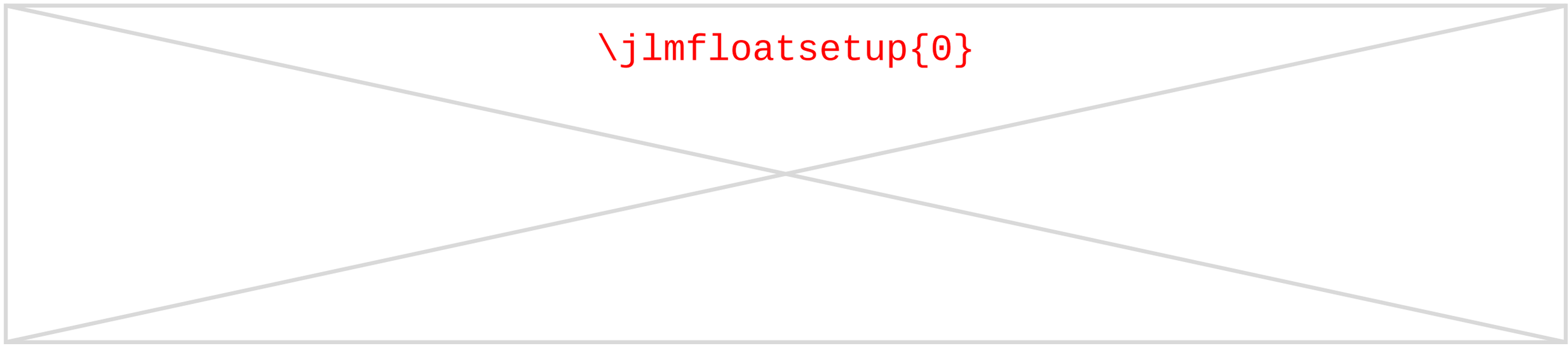


Figure 6: The division of space between a figure and a caption in layout 0. This layout should be used in case of oversized pictures or extremely long captions. This is the only layout where the caption is placed below the float. The text of the caption fits in normal text width, but the picture is allowed to extend into the space usually reserved for the caption

Figure 1: The division of space between a figure and a caption in layout 1 (default)

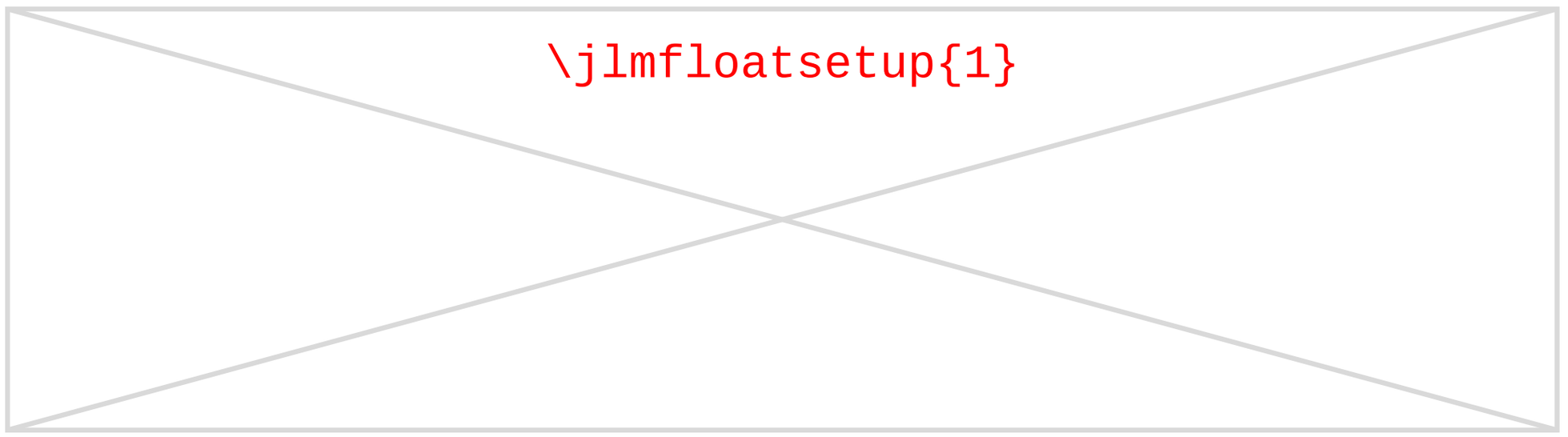


Figure 2: The division of space between a figure and

Figure 3:
The division of space between
a figure and a caption in layout 3

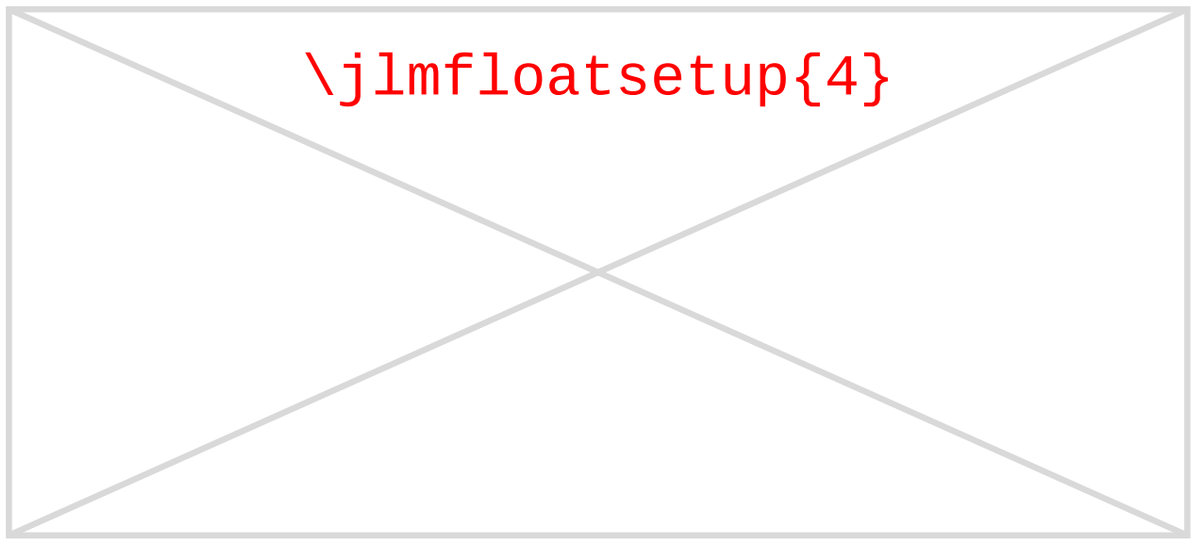
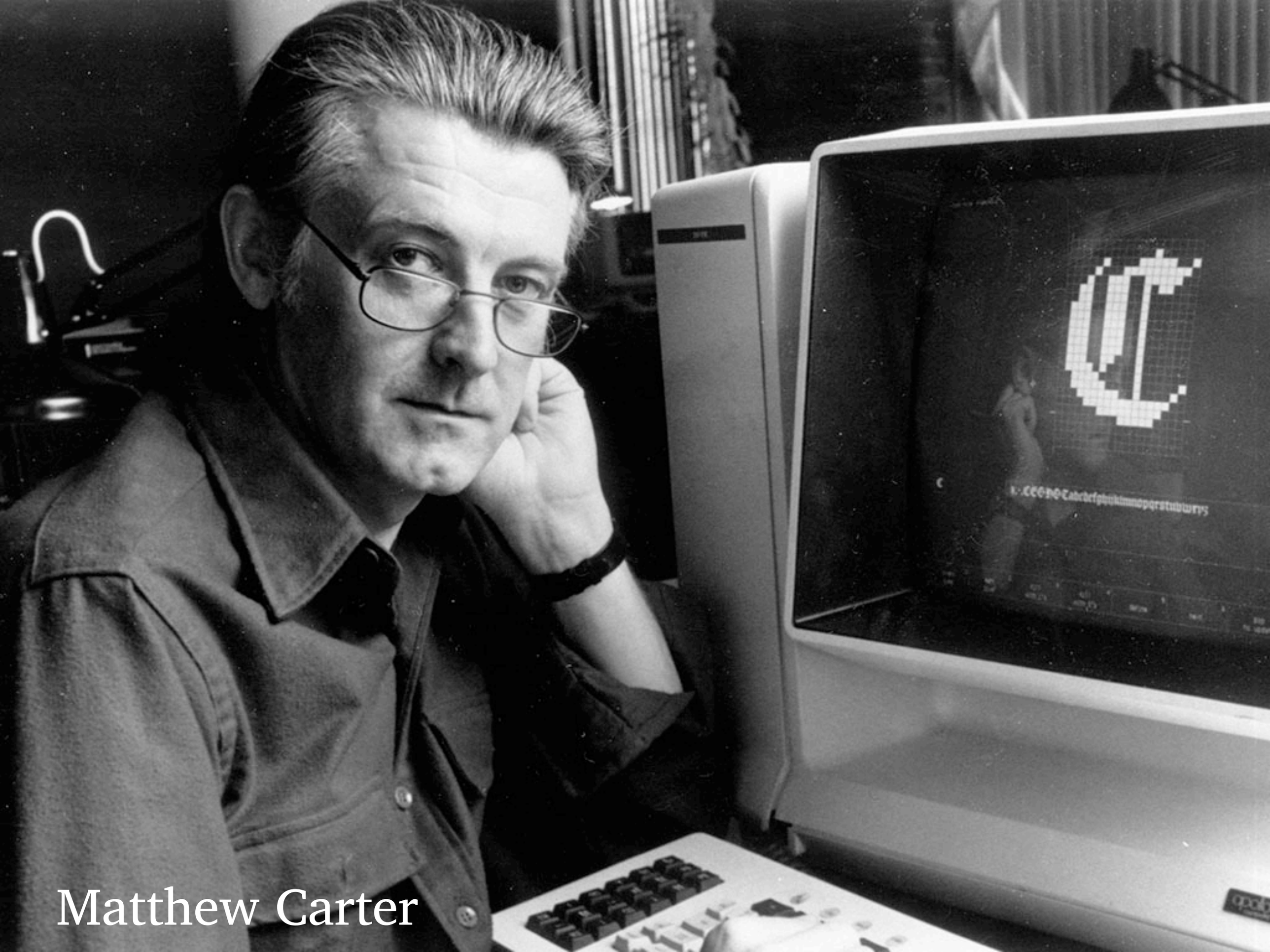


Figure 4:
The division of space between a figure and
a caption in layout 4

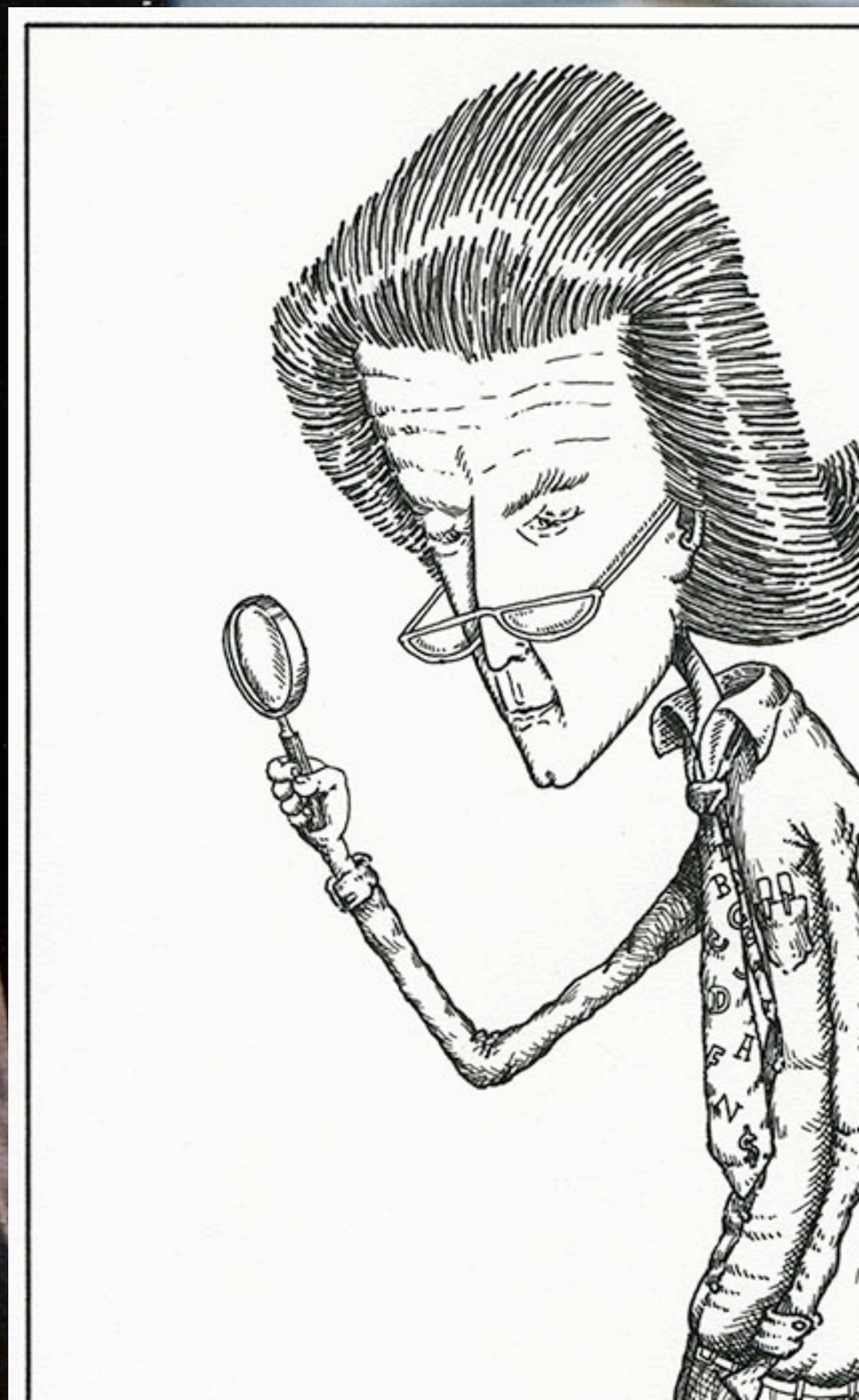
Figure 5:
The division of space between a figure and a caption in
layout 5. This is the other extreme where a very long
caption can accompany a very small picture

Fonts

```
\setmainfont[SmallCapsFeatures={LetterSpace=3},
  SizeFeatures={
    {Size={-6},Font=Charlet SL XS},
    {Size={6-8}, Font=Charlet SL S},
    {Size={8-}}}]
  {Charis SIL Compact}
\setmonofont[Scale=.92]
  {Cousine}
\setsansfont[Scale=.9,
    BoldFont={* Semibold}]
  {Open Sans}
\newfontfamily\titling@font{Playfair Display}
```



Matthew Carter



Wrigley (72pt)

Bitstream Charter®
Typeface

Ch

ABCDEFGHIJKLMNOPQ
RSTUVWXYZ&abcdefghij
klmnopqrstuvwxyzfifl123
4567890£$¢ƒ%‰✓·.,:;“”’,
„?!/()[]*†‡§-—–ÅÅÆØŒ
ååæøœßı‹›«»¿¡´ˆˇ˘˙¨˚˛˜˝
ˇˇ ˉ˘˘°° —1234567890
ˌˌ
{}ĐŁĄĄĘĘđłĽĽjąąęęĿLl·l·
·″″’’ -–ØABCD
–– · · , , ‚ ‚ ‘ ‘
EFGHIJKLMNOPQRSTUVWX
YZ&ÆŒĄĄĐĘĘIJŁĽ.,:;“”’’?!-Å
ÅPt—ųų”Þþð¥\ijIJ¶fflffiff❁
✣✣✣©℗®™@#

It should be clear, thus, that we unc
guage modelling very broadly – much
construed in speech recognition or sta
Typical articles published in JLM are e
tic generalisations – their application in
and their discovery in language corpora
from precise linguistic analyses of phon
tactic, semantic and pragmatic language

Charis SIL *by SIL International (OFL)*

Ð Ⓓ ↓ Ð Ⓓ Ç A a a ɑ ɑ A A A A ᵃ a ᵃ a ᴬ á Á á
á á Á Á Á Á à À à à à À À À À ă ă ă ă Ă Ă Ă Ă
Ă ắ ắ ắ ắ ắ ắ ắ ắ Ắ Ắ Ắ Ắ Ắ Ắ Ắ Ắ Ắ ằ ằ ằ ằ ằ
ằ ằ Ằ Ằ Ằ Ằ Ằ Ằ Ằ Ằ Ằ ẵ ẵ ẵ ẵ ẵ ẵ ẵ Ẵ Ẵ Ẵ Ẵ
Ẵ Ẵ Ẵ Ẵ Ẵ ẳ ẳ ẳ ẳ ẳ ẳ ẳ ẳ Ẳ Ẳ Ẳ Ẳ Ẳ Ẳ Ẳ Ẳ â
Â â â â Â Â Â Â á ấ ấ ấ ấ ấ ấ ấ Ấ Ấ Ấ Ấ Ấ Ấ Ấ
Ấ Ấ Ấ ầ ầ ầ ầ ầ ầ ầ Ầ Ầ Ầ Ầ Ầ Ầ Ầ Ầ Ầ ẫ ẫ
ẫ ẫ ẫ ẫ ẫ ẫ Ẫ Ẫ Ẫ Ẫ Ẫ Ẫ Ẫ Ẫ Ẫ Ẫ ẩ ẩ ẩ ẩ ẩ ẩ ẩ
ẩ Ẩ Ẩ Ẩ Ẩ Ẩ Ẩ Ẩ Ẩ Ẩ Ẩ ă ă ă ă Ă Ă Ă Ă Ă Å å ả
ả å Å Å Å Å ắ ắ ắ ắ Ắ Ắ Ắ Ắ Ắ Ä ä ä ä ä Ä Ä Ä
Ä ắ ắ Ắ ằ ằ Ằ ẵ ẵ Ẵ ặ ặ Ặ ẵ ẵ Ẵ ặ ặ ặ ặ Ặ Ặ Ặ
Ặ Ặ ã Ã ã ɑ ɑ Ã Ã Ã Ã ắ ắ Ắ ằ ằ Ằ ẵ ẵ Ẵ ȧ ȧ ȧ
ȧ Ȧ Ȧ Ȧ Ȧ Ȧ ā ā ā ā Ā Ā Ā Ā Ā ą ą Ą ą ą Ą ą ą
Ą ą ą Ą ą ą Ą ā ɑ ɑ ā Ā Ā Ā Ā Ā ắ ắ Ắ ẵ ẵ Ẵ ả
ả ả ả Ả Ả Ả Ả Ả ȧ Ȧ ằ ằ ằ ằ Ằ Ằ Ằ Ằ Ằ ẵ ẵ ẵ ẵ

Â Â Â Â Â ạ ạ Ạ ạ́ ạ́ Ạ́ ạ ạ ạ ạ Ạ Ạ Ạ Ạ Ạ ạ́ ạ́ Ạ́
ạ̀ ạ̀ Ạ̀ ặ ặ ặ ặ Ặ Ặ Ặ Ặ Ặ â̤ ậ̤ ậ̤ â̤ Â̤ Â̤ Â̤ Â̤ Â̤ ă̤ ậ̤
Ặ̆ ã̤ ậ̤ Ã̤ ạ̰ ạ̰ ạ̰ ạ̰ Ạ̰ Ạ̰ Ạ̰ Ạ̰ Ạ̰ ạ̰ ạ̰ Ạ̰ a̲ a̲ a̲ Ạ̲ á̲ ạ̲́ Á̲ à̲
ạ̲̀ Ạ̲̀ â̲ ậ̲ Â̲ ă̲ ậ̲ Ă̲ ä̲ ạ̲̈ Ä̲ æa AA AA Æ æ æ Æ Æ Æ á ǽ
Á́ Á́ Á́ ā ā Ā Ā Ā ǣ Ǣ ao AO AO å ả ả å Å Å Å Å
ą ą ą ą Ą Ą Ą Ą Ą Ą á́ ą́ Á́ ą̀ ą̀ Ą̀ A a a a a A A
A A A a a a a Æ æ æ ɐ ɐ V V V ɐ ɑ ɑ ɑ ɑ ɑ Á à
Ĉ̌ Č̌ ɑ̰ ɑ̰ ɑ̰ ɑ̰ ɒ ɒ ɒ ᗡ ᗡ ᗡ B b B b B b̃ B̃ ḃ Ḃ ḃ ḅ Ḅ
ḅ ḅ Ḅ Ḅ в ɦ ɦ Ƀ ƀ ƀ Ƀ Ƀ ʙ ʙ ƀ ɓ ɓ ɓ ɓ ɓ ɓ ɓ ɓ
ʙ Б ʙ б Б Б б C c C c c C c ċ ċ ċ ĉ Ĉ Ĉ č Č Č ċ Ċ
ċ Ç ç Ç c̄ C̄ č̣ Č̣ c̲ C̲ ć ć Ć ḉ ḉ Ḉ c ȼ Ȼ Ȼ ȼ ȼ ȼ ƈ
Ƈ Ƈ ɕ ɕ ɕ ɕ ɕ ɔ Ɔ D d D d d D d D ď ď Ď Ď Ď ḋ Ḋ
Ḋ ḍ Ḍ Ḍ ḍ Ḍ Ḍ ḍ̂ Ḍ̂ Ḍ̂ ḏ̲ Ḏ̲ Ḏ̲ đ đ Đ Đ Đ ð Đ Đ ð ð
ð̄ ð̄ ð̣ Ḍ ɗ ɗ ɗ ɗ dz ʣ Dz DZ DZ DZ dž Dž DŽ DŽ DŽ
ʤ ʤ ᴅ ᴆ ƌ ƌ ƌ ƌ ɖ ɖ ɖ ɖ Ð ɗ Ɗ Ɗ ɗ ɖ ɖ ɖ d ɑ

ɑ | ɑ̣ | δ | E | e | E | e | e | e | E | É | é | É | è | È | È | ě | Ě | Ě | ê | Ê | ế | ế | Ế | Ế
ế | ế | ề | ề | Ề | Ề | Ề | Ề | ễ | ễ | Ễ | Ễ | Ễ | Ễ | ể | ể | Ể | Ể | Ể | Ể | ě | Ě | Ě | ẻ
Ể | ë | Ë | Ë | ế | Ế | ë | Ề | ê̈ | Ê̈ | ë | Ë̃ | ẽ | Ẽ | Ẽ | ế | Ế | ề | Ề | ê̈ | Ê̈ | ė | Ė | Ė
ę | Ę | Ę | ę́ | Ę́ | ę | Ę | ę̌ | Ę̌ | Ę̌ | ę̂ | Ę̂ | ę̌ | Ę̌ | ē | Ē | Ē | é | É | É | è | È | È | ê
Ê̄ | ẻ | Ẻ | Ẻ | ė | Ė | ề | Ề | Ề | ê | Ê | Ê | ẹ | Ẹ | Ẹ | ẹ́ | Ẹ́ | ẹ | Ẹ | ệ | Ệ | Ệ | ệ | Ệ
ę̃ | Ę̃ | ę | Ę | Ę | ę | Ę | Ę | ę | Ę | ę́ | Ę́ | ę | Ę | ę̌ | Ę̌ | ë | Ë | ę́ | Ę́ | ē | È
ę́ | Ę́ | ę | ę | Ę | Ę | Ę | Ę | ę́ | Ę́ | ę | Ę | ɛ | ɛ́ | Ɛ́ | Ɛ | ɛ̃ | ɛ̃ | Ɛ̃ | Ɛ̃ | ę | ę | ę | ę
ɐ | ə | Ǝ | Ǝ | Ǝ | ə́ | Ǝ́ | ə | Ǝ | ə | Ǝ | ə̌ | Ǝ̌ | ə | Ǝ | ə̃ | Ǝ̃ | ə | Ǝ | ɘ | Ǝ | ɘ | Ə
ə | ᵊ | ₔ | ə́ | ə́ | ə | ə | ə | ə | ə̌ | ə̌ | ə | ə | ə | ə̃ | ə̃ | ə | ə | ᶏ | ᶏ | ᶏ | ᶏ | ɛ
ɛ | ɛ | ᵋ | é | É | è | È | ê | Ê | ě | Ě | ë | Ë | é | É | è | È | ẽ | Ẽ | ē | Ē | ę | Ę | ę | Ę
ë | Ë | è | È | é | É | ę | ę | ę | ɘ | ɚ | ɜ | ᵌ | ᵌ | ᵌ | ʒ | ʒ | ʒ | ʒ | ɜ | ᵌ | ɝ | ɞ | ɵ
ɷ | ɤ | ɣ | F | f | f | ff | ff | ff | F | F | FF | FF | ᶠ | f | f | f | f̃ | F̃ | ḟ | ḟ | Ḟ | Ḟ | Ḟ | ḟ
Ḟ | fi | fi | fi | ffi | ffi | ffi | FI | FI | FFI | FFI | fl | fl | fl | ffl | ffl | ffl | FL | FL
FFL | FFL | fŋ | F | ɬ | ɬ | ɬ | ɬ | ʃ | ʃ | ʃ | ʃ | ʃ | Ʃ | Ʃ | Ⱨ | G | g | g | G | G | ᵍ | ᴳ | ǵ
ǵ | Ǵ | Ǵ | Ǵ | ǧ | ǧ | Ǧ | Ǧ | Ǧ | ĝ | ĝ | Ĝ | Ĝ | Ĝ | ǧ | ǧ | Ǧ | Ǧ | Ǧ | g̈ | g̈ | G̈ | ġ

ġ Ġ Ġ Ġ ǵ ǵ Ģ Ģ Ǵ ḡ ḡ Ḡ Ḡ Ḡ ģ ģ Ģ ģ ģ Ᵹ ɠ g ᵍ
ᶢ ᶢ ᶢ ɢ ɡ ɡ ɡ ɢ ɢ ɢ ɢ ɢ ɡ ᶢ ᶢ ᶢ ʃ ʃ ʃ ʃ ɡ ɣ ɣ
ɤ ˠ ɑ̨ Ɑ̨ ɑ̨ H h H ʰ ʰ ᴴ ĥ Ĥ Ĥ ȟ Ȟ ȟ ḧ Ḧ ẗ ḣ Ḣ ḣ
ḫ Ḫ Ḫ ḥ Ḥ ḥ ḫ Ḫ Ḫ ẖ Ḫ ħ ħ Ħ Ħ ħ ħ ħ ɧ Ʉ ʉ ɧ
ɧ ḫ Ḫ Ḫ ẖ ẖ Ḫ Ḫ ɦ Ɦ Ɦ ɧ ɧ ɧ ɧ ɧ I i i 1 1 I I I I
i i I i 1 i 1 I í Í í Í Í ì Ì ì ì ì ǐ Ǐ î Î î î Î ǐ ǐ Ǐ ǐ ǐ
i̊ I̊ ï Ï ï ï ï ḯ ḯ Ḯ ḯ ḯ ì̈ Ì̈ î̈ Î̈ ĩ̈ Ĩ̈ ĩ ĩ Ĩ Ĩ Ĩ ḯ Ḯ ì̈ Ì̈ î̈ Î̈
i̇ i̧ I̧ i̱ I̱ ì̧ Ì̧ î̧ Î̧ ǐ̧ Ǐ̧ i̧ i̧ ɪ̨ I̧ I̧ I̧ I̧ i̱ I̱ ḯ Ḯ ì̧ Ì̧ ī ī Ī Ī Ī ḯ Ḯ
ì̇ Ì̇ î̇ Î̇ i̇ i̇ İ i̇ i̇ i̇ i̇ İ ì̇ ì̇ Ì̇ Ì̇ Ì̇ î̇ î̇ Î̇ Î̇ Î̇ i̧ I̧ i̱ I̱ i̧ I̧ i̧ i̧
ɪ̇ I̱ I̱ I̱ i̱ I̱ I̱ I̱ i̱ I̱ i̱ I̱ ì̱ Ì̱ î̱ Î̱ ǐ̱
Ǐ̱ ï̱ Ï̱ ī̱ Ī̱ ī̱ Ī̱ ii II iii III ij IJ IJ iv IV ix IX 1 1 I I I I
ᵻ ᵻ ᵻ I̱ ɪ̣ ᶦ i̱ ɨ Ɨ ɨ ᶧ ᶧ ᵼ ᵼ ᵼ ḯ Ḯ ì̱ Ì̱ î̱ Î̱ ǐ̱ Ǐ̱ ï̱ Ï̱ ĩ̱ Ĩ̱ ī̱ Ī̱ ĩ̧
Ɨ̃ i̱ Ɨ̱ ɨ ᵻ ᵻ ᵻ ᶧ ᵻ ᵻ ᵻ i̧ ɪ̨ i̧ i̧ i̧ ɪ ɪ ɪ ᶦ ᵻ ᵻ ᵻ ḯ Ḯ ì̱ Ì̱ î̱ Î̱
ǐ̱ Ǐ̱ ĩ̱ Ĩ̱ ɪ̧̃ Ĩ̧ ì̱ Ì̱ ī̱ Ī̱ ɨ ᵻ ᵻ ᵻ J j ɟ J ʲ ᴶ ʲ ᴶ ĵ Ĵ Ĵ ǰ j̃ J̃
j̇ J̇ ɟ ᴶ ɟ ɟ ᶡ ᶡ ɟ ɟ ɟ ɟ ᶡ ᶡ ɟ̇ ɟ̇ ɟ̇ ɟ̇ ɟ̇ ɟ ɟ ɟ ɟ ɟ ʃ ʃ K

k ᴋ ᵏ ᴷ ḱ Ḱ ḱ ǩ Ǩ ǩ k̇ K̇ ķ Ķ ķ k̄ K̄ ķ Ķ ķ k̲ K̲ ᴋ
ⱪ ⱪ ⱪ ⱪ ƙ Ƙ ᴋ ⱪ Ⱪ ⱪ ⱪ ⱪ Ⱪ Ⱪ ⱪ Ⱪ ᴋ ʞ L l l L L l
L ˡ ᴸ Í Í Ĺ Ĺ Ĺ Ǐ ľ Ľ Ǎ ľ Ǎ ļ Ļ ḷ ḷ̣ ḷ ḷ̣ ḷ̲ ḷ̣̲ ḷ̲
ḹ ḹ̣ ḹ̲ ḷ̂ ḷ̣̂ ḷ̲̂ ḹ ḹ̣ ḹ̲ l· L· L· lj Lj LJ LJ LJ ǉ Ll Ll
ls lz ɬ Ɬ Ɬ L ᴸ ɬ ɬ ɬ ɬ ɬ Ɬ Ɬ ɬ ɬ Ɬ ɬ ɬ ɬ ɬ ɬ Ɬ
Ɬ ɬ ɬ ɬ ɮ ɮ ɮ ɮ ˪ ɫ ɫ ɫ ɫ ɫ ɮ ʎ ʌ M m ᴍ m ᵐ
M ᵐ m m m m ᴹ ḿ Ḿ ḿ m̀ M̀ m̂ M̂ m̌ M̌ m̊ M̊ m̃ M̃ ṁ
Ṁ ṁ m̄ M̄ ṁ Ṁ ṃ Ṃ ṃ m̲ M̲ ᴍ ɱ m m m mʝ mʝ mʝ mʝ
mɲ Mɲ ᴍɲ Mɲ ᵐɲ m m m Ⱳ Ⱳ N n ɴ ⁿ ᴺ ǹ Ǹ ǹ n̂ N̂ ň Ň
Ň n̊ N̊ Ñ ñ Ñ ṅ Ṅ ṅ ṇ Ṇ Ṇ n̄ N̄ ṅ Ṅ ṇ Ṇ Ṇ n̂ N̂ N̂ n̲
N̲ N̲ nj Nj NJ NJ NJ ń Ń Ń ɴ ᴺ n n n и и ʁ ʁ ʁ ʁ ɲ
ɲ ɲ N ɲ N ɲ ʰɲ n n n ɲ́ Ń ɲ̄ Ń ɳ ɳ ɳ ŋ ŋ ŋ ŋ ᵑ ŋ ŋ
ŋ ŋ̊ ŋ ŋ ŋ ŋ ɴ ɲ ŋ ɒ ŋ ɲ ɴ ɒ ᵑ ŋ ŋ ŋ ŋ ŋ́ Ń́ ń́ ŋ́
ɒ́ ŋ̀ Ǹ ǹ ŋ̀ Ɒ̀ ŋ̂ N̂ n̂ ŋ̂ Ɒ̂ ŋ̃ Ñ ñ ŋ̃ Ɒ̃ ŋ̣ Ṇ ṇ ŋ̣ ɒ̣ o
o O º ᵒ º ₒ ᴼ ò Ò ò ǒ Ǒ ǒ ô Ô ô ố ố Ố Ố ố ố ồ ồ

Ồ Ồ ồ ồ õ̃ õ̃ Õ̃ Õ̃ Ỗ Ỗ ỏ̊ ỏ̊ Ố Ổ ổ̊ ổ̊ ǒ Ǒ ǒ o̊ O̊ Ö
ö ö ő́ Ő́ ȍ Ȍ ô̊ Ô̊ ö̃ Ö̃ ȫ Ȫ ȫ ő̄ Ő̄ ő̄ ȭ Ȭ ȭ ő̄ Ő̄ Ő̄ ȏ
Ỗ ỗ Ỗ ö̃ Ö̃ ö̃ ō̃ Ō̃ ō̃ o̱ O̱ o̱ ȱ Ȱ ȱ Ø ø ø ǿ Ǿ ǿ ø̲
Ø̲ ǫ Ǫ ǫ̇ Ǫ̇ ǫ̂ Ǫ̂ ǫ̌ Ǫ̌ ǫ ǫ̨ Ǫ Ǫ̨ ǫ ǫ̨ ȍ Ȍ ǭ ǭ Ǭ Ǭ ǭ
ǭ ō Ō ō ő́ Ő́ ő́ ȍ Ȍ ȍ ô̊ Ô̊ o̊ O̊ o̊ ȯ́ Ȯ́ ȏ Ȏ ȏ ô Ô
ô ơ Ơ ơ ờ Ờ ờ ỡ Ỡ ỡ ở Ở ở ợ Ợ ợ ọ Ọ ǫ̣ Ǫ̣ ọ Ọ
ọ ọ̀ Ọ̀ ộ Ộ ộ ọ̌ Ọ̌ ọ̃ Ọ̃ ǫ̃ Ǫ̃ ǭ Ǭ ȍ Ȍ ô̄ Ô̄ ǒ̄ Ǒ̄ ȫ Ȫ
ȍ Ȍ ő́ Ő́ Œ œ œ ǿ Ǿ ȍ Ȍ ô̊ Ô̊ ǒ̊ Ǒ̊ œ Œ œ ó Ó
Ó ǫ́ Ǫ́ ó Ǫ́ ớ Ớ Ở ǫ̣ Ǫ̣ ọ̣ Ọ̣ ǭ Ǭ ɔ ɔ̹ Ɔ œ æ ɒ ɔ ɔ Ɔ
Ɔ ɔ ɔ ɔ̓ ɔ̓ ɔ́ Ɔ́ ɔ̀ Ɔ̀ ɔ̂ Ɔ̂ ɔ̌ Ɔ̌ ɔ̈ Ɔ̈ ɔ̋ Ɔ̋ ɔ̏ Ɔ̏ ɔ̃ Ɔ̃ ɔ̏ Ɔ̏ ɔ̄
Ɔ̄ ɔ̰ Ɔ̰ ɔ̱ Ɔ̱ ɔ̈ Ɔ̈ ɔ̀ Ɔ̀ ɔ́ Ɔ́ ɔ ɔ ɔ̜ ɔ̰ ɔ̰ ɔ̰ ɔ̰ ɔ̰ ɔ̜ ɔ̜ ʊ ʊ ɵ
ɵ Ɵ ɵ ɵ ɵ ɵ ɵ ǿ Ǿ ɷ ȣ ȣ ȣ ȣ ȣ ȣ ȣ ȣ ȣ P p P p P
ṕ Ṕ Ṕ p̃ P̃ ṗ Ṗ Ṗ p̄ P̄ p̱ P̱ ᴘ ƥ Ƥ ƥ ƥ ƥ ƥ ƥ ƥ ƥ ƥ ƥ
ƥ ƥ ƥ ƥ ƥ ƥ ƥ Ƥ Ƥ Ƥ ɋ ɸ ɸ ɸ ɸ ɸ ɷ Q ɋ Q ɸ ɸ ɸ
ɸ ɸ ɋ ɋ ɋ ɋ ɋ ɋ ᴋ R ʀ R r r r r R ŕ Ŕ ŕ ř Ř

ř ñ Ñ ṙ Ṙ ṙ ṛ Ṛ Ṛ ȓ Ȑ Ȑ r̂ R̂ R̂ ṛ Ṛ Ṛ r̄ R̄ R̄ r̥ R̥ r̲
R̲ R̲ ʀ ʀ R ʁ ɾ ʀ ʀ ɾ ɾ ɾ ɾ ɾ ɹ ɹ ʁ ɻ ɾ ɾ ɾ ɾ
ɻ ɻ ɻ ɾ ɾ ɾ ʀ ɾ ʀ ɾ ɾ ɾ ɾ ɾ ɾ ɾ ɾ ɾ ʁ ʁ S s s ˢ ŝ Ŝ
ŝ š Š š š Š š ṡ Ṡ ṡ ṣ ş Ṣ Ṣ ş ṣ ṣ Ṣ ṣ ṣ̌ Ṣ̌ ṩ Ṩ ṩ ṣ
Ṣ ṣ ſ ı̇ ß ß ś ś Ś ś́ ś́ Ś́ s s s s s ʂ ʂ ʂ ʂ ʂ ʂ ʂ ʂ ʂ
ʃ ʃ ʃ Σ Σ ˢ ʃ ʃ ʃ ʃ ʃ ʃ ʃ ɳ ɭ ʃ ʃ ʃ ʃ T t T
t t T ť t̓ Ť Ť Ť ẗ ṫ Ṫ Ṫ ṭ ţ Ṭ Ṭ Ṭ Ṭ t̄ T̄ ṭ Ṭ Ṭ ṭ Ṭ Ṭ
t̪ T̪ T̪ t̲ T̲ T̲ tɕ tʰ ts ʒ tʃ ʧ ʧ ʇ ŧ Ŧ Ŧ ƚ Ƭ ƭ ŧ ŧ ŧ
ŧ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʈ ʉ U u ᴜ ᵘ ᵘ ᵘ ᵁ ú Ú
ú ù Ù ù ũ Ũ ŭ Ŭ ŭ û Û û ǔ Ǔ ǔ ů Ů ů Ü ü ü ű Ű
ű ù̃ Ù̃ ù̃ û̃ Û̃ ŭ̃ Ŭ̃ ŭ̃ ũ̃ Ũ̃ ũ̃ ū̃ Ū̃ ū̃ ű̃ Ű̃ ű̃ ũ̃ Ũ̃ ũ̃ ű̃
ű̃ ů̃ ů̃ Ů̃ ů̃ ų̃ Ų̃ ų́̃ Ų́̃ ų̀̃ Ų̀̃ ų̂̃ Ų̂̃ ų̌̃ Ų̌̃ ų̃ ų̃ Ų̃ Ų̃ ų̃ ų̃ ū
Ū ū ű́ Ű́ ù̇ Ù̇ û̇ Û̇ ü̇ Ü̇ ü̇ ů̇ Ů̇ ů̇ ų́̇ Ų́̇ ų̀̇ Ų̀̇ ų̂̇ Ų̂̇ ư
Ư ư ứ Ứ ứ ừ Ừ ừ ữ Ữ ữ ử Ử ử ự Ự ự ų̇ Ų̇ ų́̇ Ų́̇
ų̇ Ų̇ ụ Ụ ụ ú́ Ú́ ụ̀ Ụ̀ ụ̂ Ụ̂ ụ̌ Ụ̌ ụ̃ Ụ̃ ų̇ Ų̇ ụ̇ ų̇ Ų̇ ụ̇ ų̇ Ų̇

Ꭹ	ᶣ	Ꭹ	ᶣ́	Ꭹ́	ᶣ̀	Ꭹ̀	ᶣ̂	Ꭹ̂	ᶣ̌	Ꭹ̌	ᶣ̈	Ꭹ̈	ᶣ̋	Ꭹ̋	ᴜ	ᵁ	ᵘ	ᵘ	ᵘ	ᴝ	ᴞ	ᶣ
æ	ʉ	Ʉ	ᵾ	Ʉ̄	ᵾ̄	ʰ	ʰ̄	ʰ̄	ʰ̄	ʉ́	Ʉ́	ʉ̀	Ʉ̀	ʉ̂	Ʉ̂	ʉ̌	Ʉ̌	ʉ̈	Ʉ̈	ᶣ̄	Ꭹ̄	ᵾ
ʊ	ʊ	ʊ	ᶣ	ᶣ	ᶣ	ᶣ	ᵡ	ᵡ	ᵡ	ᵡ	ᵡ	ᵡ	ɯ	Ɯ	ɯ	ʷ	Ɯ	ᵡ	ᵡ	ᵡ	ᵡ	ᵡ
ᵾ	ᵾ	ᴝ	Ʉ	ᴝ	ᵘ	ᵘ	ᵘ	ʉ́	Ʉ́	ʉ̀	Ʉ̀	ʉ̂	Ʉ̂	ʉ̌	Ʉ̌	ʉ̄	ʊ	ʊ	ʊ	V	v	v
v	v	v	ᵛ	V	ᵛ	ᵥ	ᵛ	ṽ	ṽ	Ṽ	ṽ	ṽ	ᵽ	ᵽ	Ṽ	ᵽ	ᵽ	vi	VI	vii	VII
viii	VIII	ᴠ	ᶣ	ᶣ	ᶣ	ᶣ	ᴝ	Ʉ	ᴠ	ᴠ	Ʌ	Ʌ	ᴝ	ᴠ	ᴠ	ᵛ	ᵛ	ᵛ	ᵛ	ʉ́	Ʉ́	ʉ̀
Ʉ̀	ʉ̂	Ʉ̂	ʉ̃	Ʉ̃	ᶣ̃	Ꭹ̃	ᴠ	ᴠ	ᵠ	ᵠ	ᵠ	ɷ	ɷ	ɷ	ʌ	Λ	ʌ	Λ	Λ	ᴧ	ᴧ	ᴧ	ᴧ
W	w	w	ʷ	ᵂ	ẃ	Ẃ	ẃ	ẁ	Ẁ	ẁ	ŵ	Ŵ	ŵ	ẘ	ẅ	Ẅ	ẅ	ẇ	Ẇ
ẇ	ᶣ	Ꭹ	ᶣ	ᶣ	ᶣ	ᶣ	ᵡ	ᵡ	ᵡ	ᵡ	ᵡ	ᵡ	ʍ	X	x	x	x	ˣ	X
ˣ	ₓ	ẍ	Ẍ	ẍ	ẋ	Ẋ	ẋ	x̲	X̲	xi	XI	xii	XII	xᶣ	ᵡ	ᵡ	ᵡ	Y	y	Y	ʸ	ý
Ý	ý	ỳ	Ỳ	ỳ	ŷ	Ŷ	ŷ	ẙ	ÿ	Ÿ	ÿ	ỹ	Ỹ	ỹ	ẏ	Ẏ	ẏ	ȳ	Ȳ	ȳ	ẙ	Ẏ
Ẏ	ẏ	Ꭹ	Ꭹ	ʏ	ʏ	Ɏ	Ɏ	ᵡ	ᵡ	ᵡ	ᵡ	ʏ	ʏ	Ɣ	Ɣ	Ɣ	Ɣ	ɣ	ɣ	ɣ	Z	z	z
z	z	ᶻ	ᶻ	ᶻ	ᶻ	ž	Ž	ẑ	ẑ	Ẑ	ẑ	ẑ	ž	ž	Ž	ž	ž	ẽ	Ẽ	ẓ	ẓ	Ẓ	ẓ	ẓ
ẓ̌	Ẓ̌	ẓ̌	Ẓ̌	z̲	z̲	Z̲	z̲	z̲	ʂ	ź	ź	ź	ź	Ź	ż	ż	ż	ż	Ż	z	ᶻ	ᶻ	ᶻ	ᶻ
ᶻ	ᶻ	ᶻ	ʐ	ʐ	ᶻ	ᶻ	ᶻ	ᶻ	Ƶ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ

ẇ	ẉ	Ẉ	ẉ	ẘ	Ẉ	ẘ	Ẉ	w	ᴡ	Ᵽ	ᴡ	ᵂ	Ᵽ	ʍ	X	x	x	x	ˣ	X
ˣ	ₓ	ẍ	Ẍ	ẍ	ẋ	Ẋ	ẋ	x̲	X̲	xi	XI	xii	XII	xᶣ	ᵡ	ᵡ	ᵡ	Y	y	Y	ʸ	ý
Ý	ý	ỳ	Ỳ	ỳ	ŷ	Ŷ	ŷ	ẙ	ÿ	Ÿ	ÿ	ỹ	Ỹ	ỹ	ẏ	Ẏ	ẏ	ȳ	Ȳ	ȳ	ẙ	Ẏ
Ẏ	ẏ	Ꭹ	Ꭹ	ʏ	ʏ	Ɏ	Ɏ	ᵡ	ᵡ	ᵡ	ᵡ	ʏ	ʏ	Ɣ	Ɣ	Ɣ	Ɣ	ɣ	ɣ	ɣ	Z	z	z
z	z	ᶻ	ᶻ	ᶻ	ᶻ	ž	Ž	ẑ	ẑ	Ẑ	ẑ	ẑ	ž	ž	Ž	ž	ž	ẽ	Ẽ	ẓ	ẓ	Ẓ	ẓ	ẓ
ẓ̌	Ẓ̌	ẓ̌	Ẓ̌	z̲	z̲	Z̲	z̲	z̲	ʂ	ź	ź	ź	ź	Ź	ż	ż	ż	ż	Ż	z	ᶻ	ᶻ	ᶻ	ᶻ
ᶻ	ᶻ	ᶻ	ʐ	ʐ	ᶻ	ᶻ	ᶻ	ᶻ	Ƶ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ	ᶻ

Ʒ ʒ ʒ ʒ Ʒ Ʒ Ʒ ʒ ʒ Ʒ ʒ ʒ ᶾ ᶾ ᶾ ᶾ ǯ Ǯ ǯ ʒ ɛ Ɛ Ɛ ʒ ʒ
ʒ ʒ ʒ ʒ ʒ ʒ ʒ ʒ þ Þ Þ þ Þ þ ƨ ɛ ɛ ɛ ɛ ɬ ɬ ɬ ɬ ɬ
ɬ ɬ ɾ ɾ ɾ ƽ ƽ ƽ ɂ ɂ ʔ ʔ ʔ ʔ ʔ ʔ ʼn ʒ ʒ ' ' ' ' ' '
ˁ ˁ ˁ ˁ ʕ ˤ ʢ ʡ ʜ ᶜ ᶜ ʡ ʢ ʖ ǂ ! ƈ ⊙ ʷ ʷ ⊐ Ê

---

a A A ă Ă Ă ä Ä Ä ə Ə ə ə Ə ə æ Æ Æ б Б Б в В
в г Г Г ғ Ғ Ғ ғ ғ Ғ ғ ғ ғ ғ ғ ғ ғ Ғ Ғ Ғ ҕ ҕ
ҕ ҕ ҕ ҕ д Д Д d d ԁ ђ Ђ ђ ԁ ԁ ԁ ѓ Ѓ ѓ ʒ Ʒ ʒ e
Е е è È È ё Ё Ё ĕ Ĕ Ĕ є Є Є ж Ж ж ӂ Ӂ ӂ ӝ Ӝ ӝ
ж Ж Ж з З з ҙ Ҙ ҙ ɛ Ɛ ɛ ҽ ҽ Ҽ Ҽ ӟ Ӟ ӟ s S s Ʒ
ʒ ʒ Ʒ ʒ ʒ ҙ Ҙ ҙ и И и ѝ Ѝ ѝ ӣ Ӣ ӣ ӥ Ӥ ӥ ӣ Ӣ ӣ
і ı I I ї Ї Ї й Й Й j ȷ J ȷ к К к қ Қ қ ҟ Ҟ ҟ ҡ
Ҡ ҡ Ҡ ҡ ҡ Ҡ ҡ ҡ Ҡ ҡ q Q Q л Л л л ӆ Ӆ ӆ ӆ Ӆ ӆ
ӆ ӆ Ӆ Ӆ љ Љ Љ љ Љ љ ӆ Ӆ ӆ х Х х м М м ӎ
Ӎ ӎ н Н н ᴴ ӈ Ӈ ӈ ң Ң ң Ӈ ӈ ӊ Ӊ ӊ ҥ Ҥ ҥ

ҥ	Ҥ	ҥ	ҥ	Ҥ	ҥ	о	О	о	ӧ	Ӧ	ӧ	ѳ	Ѳ	ѳ	ё	Ё	ё	п	П	п
ҧ	Ҧ	ҧ	р	Р	ᴘ	ҏ	Ҏ	ᴘ	ᵽ	Ᵽ	ᵽ	с	С	с	ɢ	Ԍ	ɢ	ç	Ç	ç	т	Т
т	ҭ	Ҭ	ҭ	ҭ	Ҭ	Ҭ	ћ	Ћ	ћ	ќ	Ќ	ќ	у	У	у	ӯ	Ӯ	Ӯ	ў	Ў	Ў	ӱ
Ӱ	ӱ	ӳ	Ӳ	Ӳ	ү	Ү	ү	ұ	Ұ	ұ	ф	Ф	ф	х	Х	х	ҳ	Ҳ	ҳ	𝘅	𝘅	𝘟
𝙭	𝔁	𝔵	𝘅	𝘟	𝘅	𝙭	𝙭	ҳ	Ҳ	Ҳ	h	ћ	h	h	h	ц	Ц	ц	ҵ	Ҵ	ҵ	ч
Ч	ч	ӵ	Ӵ	ӵ	ҷ	Ҷ	ҷ	ҷ	Ҷ	ҷ	ҵ	Ҵ	ҵ	є	Є	є	ҽ	Ҽ	ҽ	ҵ	Ҵ	ҵ
ш	Ш	ш	щ	Щ	щ	ъ	Ъ	ъ	ы	Ы	ы	ӹ	Ӹ	ӹ	ь	Ь	ь	ҍ	Ҍ	ҍ	ѣ	
Ѣ	ѣ	э	э	Э	Э	э	э	ӭ	Ӭ	Ӭ	ю	Ю	ю	Я	я	я	ӕ	Ӕ	ӕ	ѳ	Ѳ	

ϴ	ϒ	ν	ϒ	ϖ	ϖ	ϖ	ѡ	Ѡ	ѡ	І	І	ı

α	β	ᵝ	ᵦ	Γ	γ	ᵞ	ᵧ	г	δ	Δ	ᵟ	θ	Θ	ᶿ	𝛝	𝛝	𝛝	λ	Λ	π	Π	п	ρ	ᵨ
ᴘ	σ	ᵠ	φ	χ	ᵡ	χ	Ψ	Ψ	Ω	Ω														

0	1	2	3	4	5	6	7	8	9	LRE	LRM	RLE	RLM	WJ	ZWNBSP	ZWSP							
										_	═	-	SHY	¡	¿	·	‥	…	(	)	[	]	{	}	〈	〉	§	¶

© ® @ * / / \ & # % ‰ † ‡ • ▶ · ‴ ‵ ‶ ‷ ˆ ˘ ˂ ˍ

ˍ ˍ ˂ ˍ ˍ ˍ ˂ ˍ ˍ ˍ ˂ ˍ

⅓	¼	⅕	⅙	⅛	²	₂	⅔	⅖	³	₃	¾	⅗	⅜	⁴	₄	⅘	⁵	₅	⅚	⅝	⁶
₆	⁷	₇	⅞	⁸	₈	⁹	₉	№	Rs	™	µ	LRO	PDF	RLO						ĭ	i	i	˚	~
˛																								
ˋ	ˉ																							
		.	·	·			!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	0	1	2	3	4	5	6	7	8	9	:
;	<	=	>	?	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Y	Z	[	\	]	^	_	`	f		F ( )	′	X				^	(	)	′	″	‴	^	↕	↖
↗	↘	↙	⇐	⇑	⇓	⇕	‖	◊	⟦	⟧	𝔪	𝔭	'	‘	’	,	‚	‹	›	"	“	”	„
‟	«	»	Ǐ	Ǐ	i	i	,	;	:	!	‼	?	.	·	Ԥ	ԥ	Ң	ң	≋	Ꭰ	Ч	ƚ	¬	⌐		

¨	æ	ɘ	ɞ	ø	œ	Œ	ʏ	ɤ	ħ	ʉ	ʎ																				
																															Đ

VS 8
VS 9
VS 10
VS 11
VS 12
VS 13
VS 14
VS 15
VS 16
ZW J
ZW NJ
VS 1
VS 2
VS 3
VS 4
VS 5
VS 6
VS 7

JLM typography

Main leading: 1 ygu
Main body font size: 0.75 ygu (~10.204 pt)
Indents etc.: 1.5 xgu (~20.409 pt)

Small-text font size: 0.63 ygu (~8.57 pt)
Small-text leading: 0.75 ygu (~10.204 pt)

Monospace font size: 0.69 ygu (~9.388 pt
Monospace font size: 0.58 ygu (~7.89 pt)

Tiny font size: 0.525 ygu (~7.143 pt)
Sanserif font size is: 0.675 ygu (~9.184 pt)

JLM typography

Heading Level 1: Text centered, number left-aligned, Titling font (Playfair), All small caps (“smcp” + ”c2sc” features enabled), Size: 0.9 ygu (~12.245 pt), Baseline shift: –0.6 ygu (~ –8.16pt), Space before and after paragraph: 1 ygu, Leading: 1 ygu, Tracking: +40 (4% design size)

Heading Level 2: Text centered, number left-aligned, Body font (Charis), Style: Italic, Baseline shift: 0.3 ygu (~4.08 pt), Space before: 1 ygu, Space after: none

Heading Level 3: Text centered, number left-aligned, Body font, Size: small, Leading: main, Baseline shift: 0.3 ygu, Space before: 1 ygu, Space after: none

Inside JLM vol. 0 issue 1

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

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

Keywords: corpus design, Bulgarian National Corpus, computational linguistics

tics, 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

— features freely defined by the user on the level of multi-word units, such as pragmatic labels described below (seehEXTRACATEGORIESi),
— features implemented in*Multiflex*in order to handle morphographical problems, such as letter case, initialisms and acronyms (seehGRAPHICALCATEGORIESi).

3.2.2 Pragmatic Variants

Among many variants of a name, we want to distinguish a few important pragmatic variants marked by specific values of the*Usage*category:
— the official variant (*offic*) used in official lists and documents,
— the neutral variant (*neut*) preferred in text generation,
— the neutral spoken variant (*spok*) preferred for speech generation.
In Fig. 3 the top-most path describes the elliptical variant*Rodowicz*annotated as the neutral spoken (*spok*), neutral (*neut*), or unmarked (h*E*i) variant. The bottommost path allows us to obtain the elliptical variant*„Anoda" Rodowicz*described as unmarked. In order to compactly represent several identical forms with different feature values, a feature structure can contain alternative values. Here, the alternative operator '—' allows us to reduce the graph's size from 26 to 15 paths. Note that two different values of the same category cannot be selected on the same path. Here for instance the form*Jan Anoda Rodowicz*can be generated with the unmarked value of*Usage*but not with*offic*because the path omitting quotes (i.e. components*$5*and*$7*) has the constrainth*Usage* = h*E*ii.

3.2.3 Initials and Letter Case

Urban proper names frequently take abbreviated forms of their components when appearing in written texts. Any first name can be reduced to its one or two initial letters followed by a dot as in example (1). Similar behavior can be noted in the words 'Street', 'Square', 'Avenue', as well as titles and functions such as 'General'.

Using a dictionary of abbreviations would be most appropriate, unfortunately we are not aware of such a dictionary for Polish. Thus we propose the following partial solution. Whenever an abbreviation is constructed from one to five initial letters, whether or not followed by a dot, the category *Init* (mentioned in Fig. 5) with the corresponding value can be used. For instance in Fig. 3 component*$1* (*Jan*) can be replaced by its initial (*J.*) due to the equation *Init* = *dot*.

Note that some words are abbreviated differently than by a prefix, as in *płk* for *pułkownik* 'colonel'. Moreover abbreviations or acronyms may be formed from inflected words as*W-wie*for*Warszawie*, or may become independent inflecting lexemes, e.g.*ONZ*,*ONZ-u*for*Organizacja Narod w Zjednoczonych*'United Nations Organization'. Such cases are currently described with specific dedicated inflectional graphs.

Many urban names contain capitalized common words, as for instance *ulica* ***Długa*** 'Long Str.' or ***Most*** *Syreny* 'Siren Bridge' These words are described in Morfeusz in lower case only. Thus, when such components are morphologically analyzed they obtain lower case lemmas. For instance in Fig. 1 the nickname*Anoda* is assigned a common word lemma*anoda*'anode'. In order to express this difference in spelling between the lemma and its form we have introduced the*LetterCase* category (see Fig. 6) mentioned in Fig. 5. It indicates how to transform the letter case of the lemma into the form desired in the dictionary. If no transformation is needed the value is*same*.

The*LetterCase*value is most often implicit, i.e. it does not appear in inflection graphs but is automatically deduced from each component of a compound during its morphological analysis. It can however also be explicitly used in graphs if needed.

4 GRAPH MANAGEMENT

Graphs are created and modified using an enhanced graph editor based on*Unitex* Paumier (2003). It allows for the creation, connection, filling out and deletion of boxes. A graph can be assigned to one or many entries at a time whenever they are simultaneously selected from the list of names.

As the number of graphs grows with the number of names described, managing graphs becomes difficult. Currently we have over 8,900 names with 451 corresponding graphs. The majority of names use only a few graphs, which are thus easy to remember. For the rest, however, the user needs some support.

4.1 *Filtering Graphs*

When a lexicographer introduces a new proper name,*Toposław*displays the list of currently defined graphs which have the same number of components as the name in question. This significantly reduces the number of graphs that have to be considered.

a compound during its morphological analysis. It can however also be explicitly used in graphs if needed.

4 GRAPH MANAGEMENT

Graphs are created and modified using an enhanced graph editor based on*Unitex* Paumier (2003). It allows for the creation, connection, filling out and deletion of boxes. A graph can be assigned to one or many entries at a time whenever they are simultaneously selected from the list of names.

As the number of graphs grows with the number of names described, managing graphs becomes difficult. Currently we have over 8,900 names with 451 corresponding graphs. The majority of names use only a few graphs, which are thus easy to remember. For the rest, however, the user needs some support.

4.1 *Filtering Graphs*

When a lexicographer introduces a new proper name,*Toposław*displays the list of currently defined graphs which have the same number of components as the name in question. This significantly reduces the number of graphs that have to be considered.

3.2.2 Pragmatic Variants

Among many variants of a name, we want to distinguish a few important pragmatic variants marked by specific values of the*Usage*category:

— the official variant (*offic*) used in official lists and documents,
— the neutral variant (*neut*) preferred in text generation,
— the neutral spoken variant (*spok*) preferred for speech generation.

In Fig. 3 the top-most path describes the elliptical variant*Rodowicz*annotated as the neutral spoken (*spok*), neutral (*neut*), or unmarked (h*E*i) variant. The bottommost path allows us to obtain the elliptical variant*„Anoda" Rodowicz*described as unmarked. In order to compactly represent several identical forms with different feature values, a feature structure can contain alternative values. Here, the alternative operator '—' allows us to reduce the graph's size from 26 to 15 paths. Note that two different values of the same category cannot be selected on the same path. Here for instance the form*Jan Anoda Rodowicz*can be generated with the unmarked value of*Usage*but not with*offic*because the path omitting quotes (i.e. components*$5*and*$7*) has the constrainth*Usage* = h*E*ii.

3.2.3 Initials and Letter Case

Urban proper names frequently take abbreviated forms of their components when appearing in written texts. Any first name can be reduced to its one or two initial letters followed by a dot as in example (1). Similar behavior can be noted in the words 'Street', 'Square', 'Avenue', as well as titles and functions such as 'General'.

Using a dictionary of abbreviations would be most appropriate, un-

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

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

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 inappropriate 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.

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-

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.

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

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

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 (sub-

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; adjec-

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 Tufiș (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.

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 Tufiș (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.

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 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;	Providing information, news and commentary	Relatively stable form and structure, some emphatic elements

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

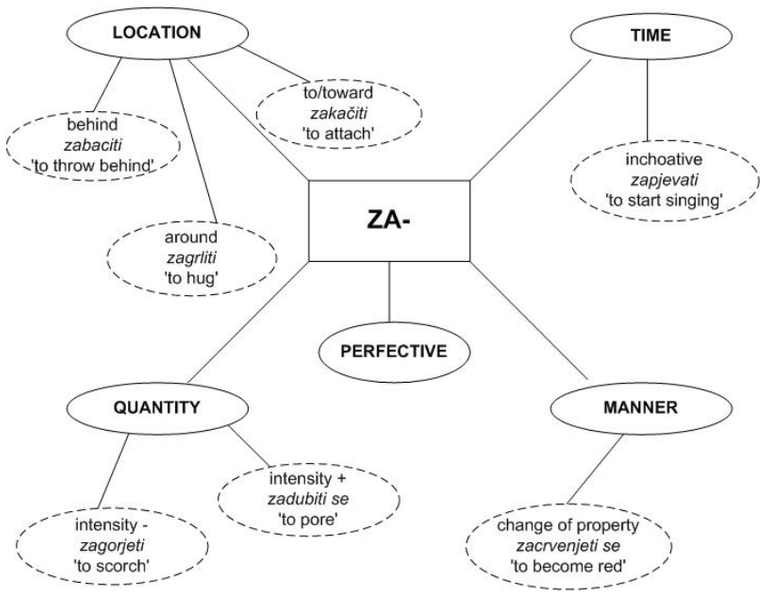
Table 2: Distribution of domains and genres across styles in the BulNC

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

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

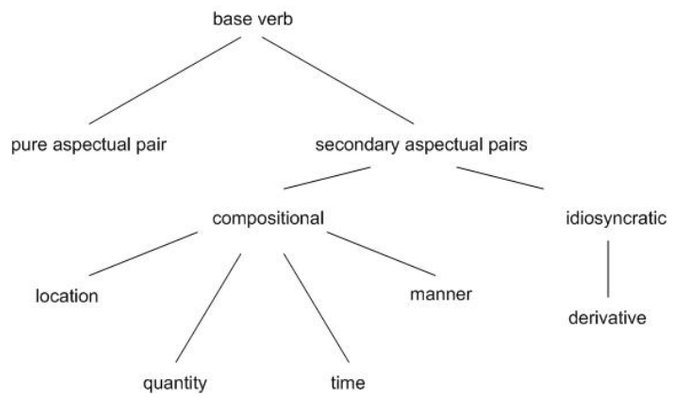


Figure 2: Derivationally motivated relations between base verbs and derivatives

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.

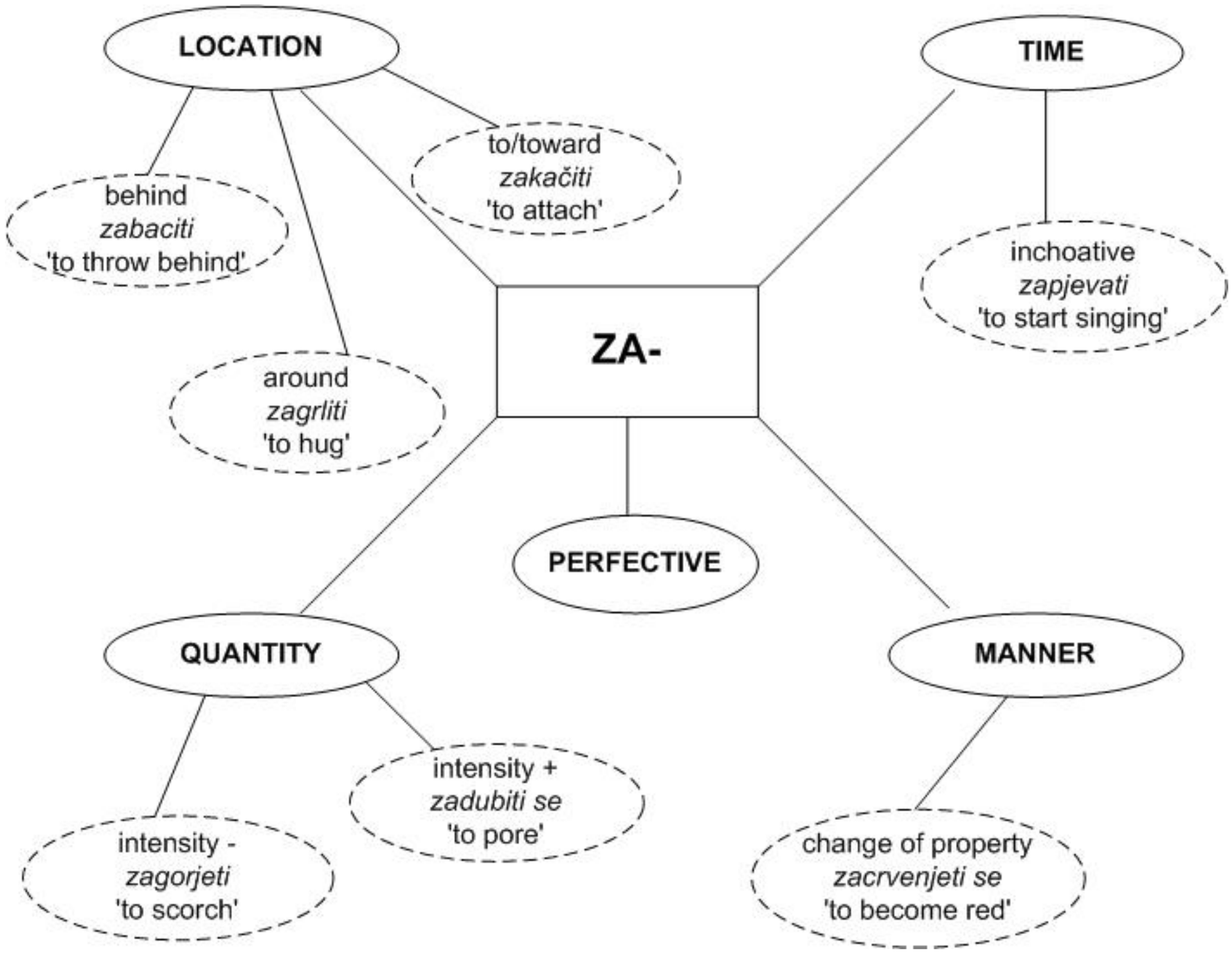
To fulfill these tasks, it is necessary to determine which prefixes take part in the derivation of particular base forms. The data on the derivational spans of verbs so far have not been systematically and extensively presented in Croatian morphology. In other words, large-scale data indicating which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

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 pre-

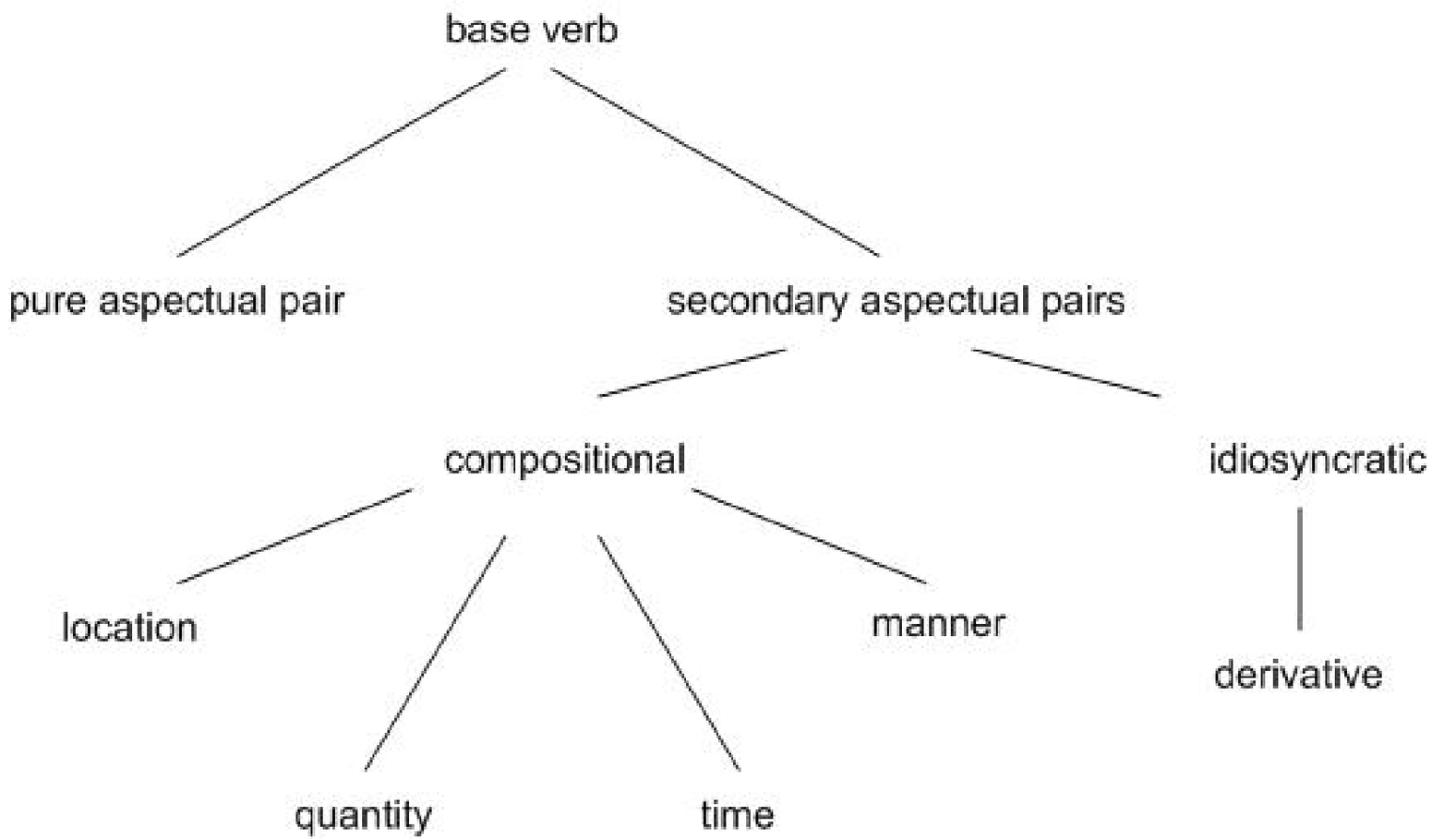


Figure 2:
Derivationally motivated relations between base verbs and derivatives

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.

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.

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	1. **inchoativity** – *zatrčati se* 'to start running', *zapjevati*	1. **intensity** – *zadubiti se* 'to pore', *zagorjeti* 'to scorch'	1. **change of property** – *zacrveniti se* 'to become red'

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-

Table 2: Morphosemantic relations in *location* group

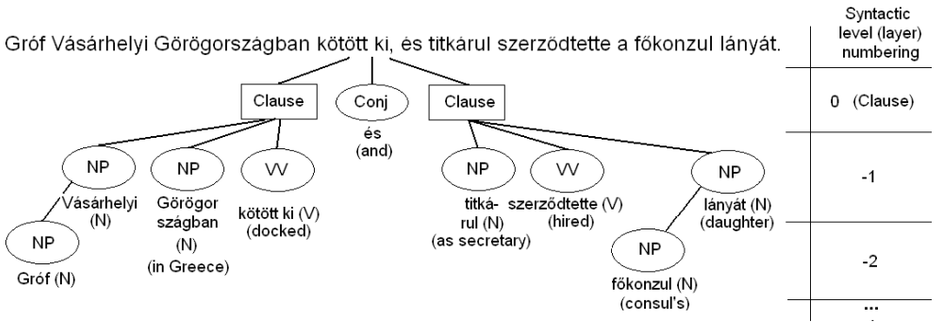


Figure 1: An example of syntactic phrase structure of the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as his secretary*)

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 *Prosody 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-

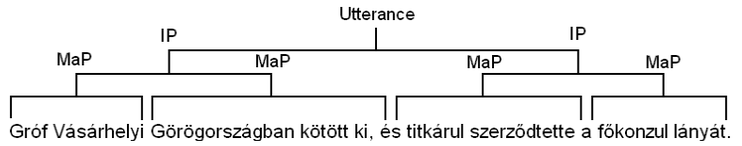


Figure 2: An example of canonical prosodic structure of a Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*"

erarchic view of the prosodic structure as follows top-down: utterances are composed of *intonational phrases* (IP), which can be divided into *phonological phrases* (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

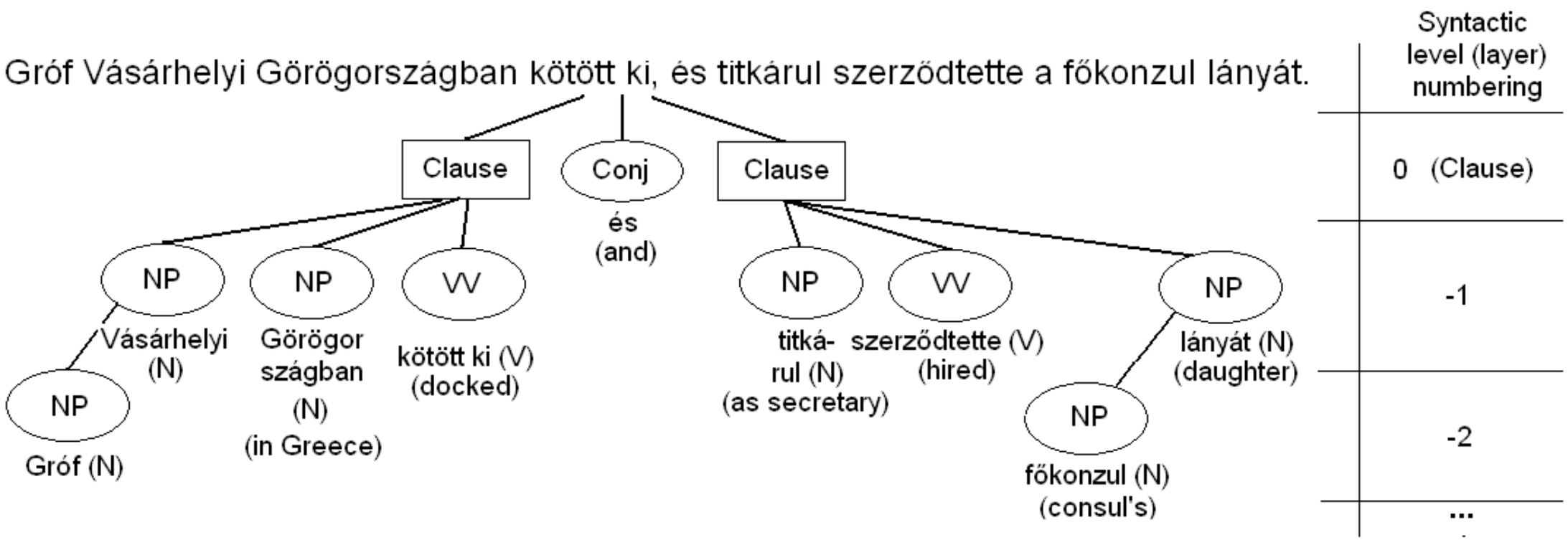


Figure 1: An example of syntactic phrase structure of the Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*" (*Count Vásárhelyi docked in Greece, and hired the daughter of the consul as his secretary*)

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

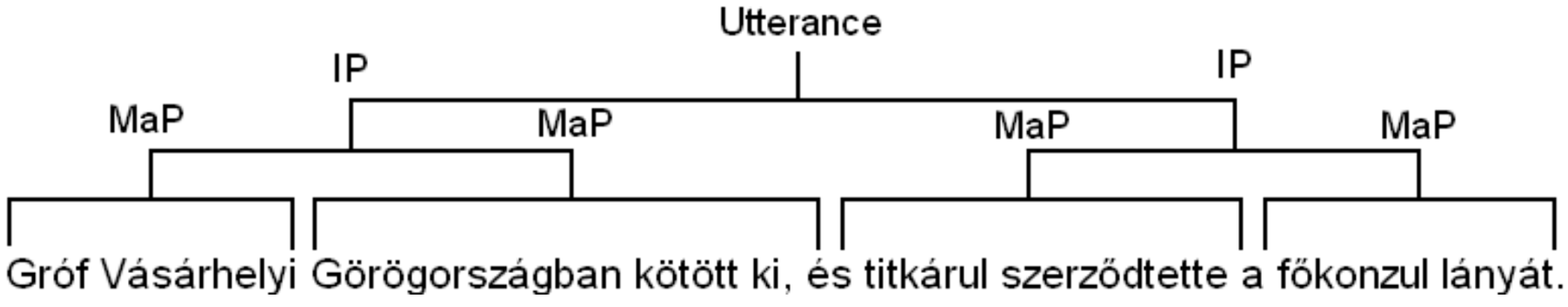


Figure 2: An example of canonical prosodic structure of a Hungarian sentence "*Gróf Vásárhelyi Görögországban kötött ki, és titkárul szerződtette a főkonzul lányát*"

erarchic view of the prosodic structure as follows top-down: utterances are composed of *intonational phrases* (IP), which can be divided into *phonological phrases* (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 re-

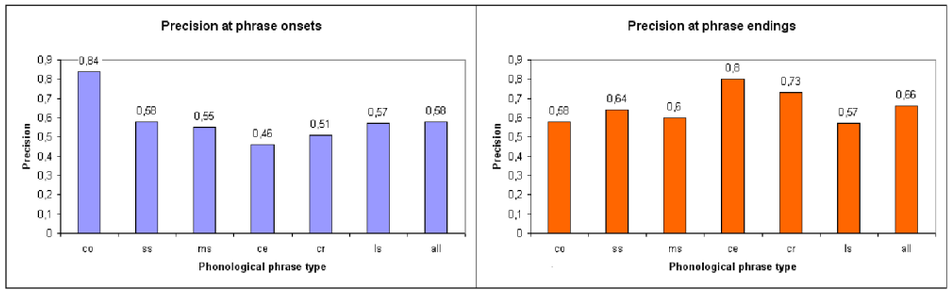


Figure 5: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right)

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 phonological 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

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)

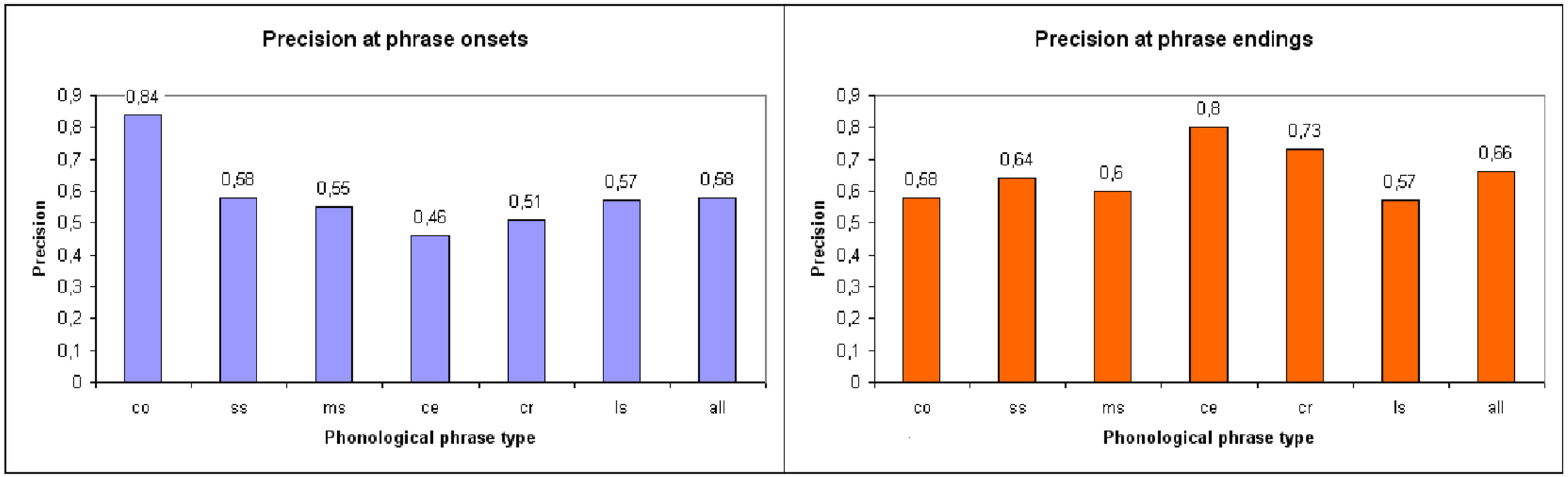


Figure 5: Precision of syntactic phrase recovery based on phonological phrase boundary detection (within 150 ms) for phrase onsets (left) and endings (right)

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

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

Table 4:
Distribution of syntactic phrase (XP) levels (or layers) based on phonological phrase type (phonological phrase endings compared to syntactic phrase endings)

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:

$$\sigma_t = \frac{1}{tp} \sum_{i=1}^{tp} |t_i - t_i^{ref}|, \tag{3}$$

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-

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

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.

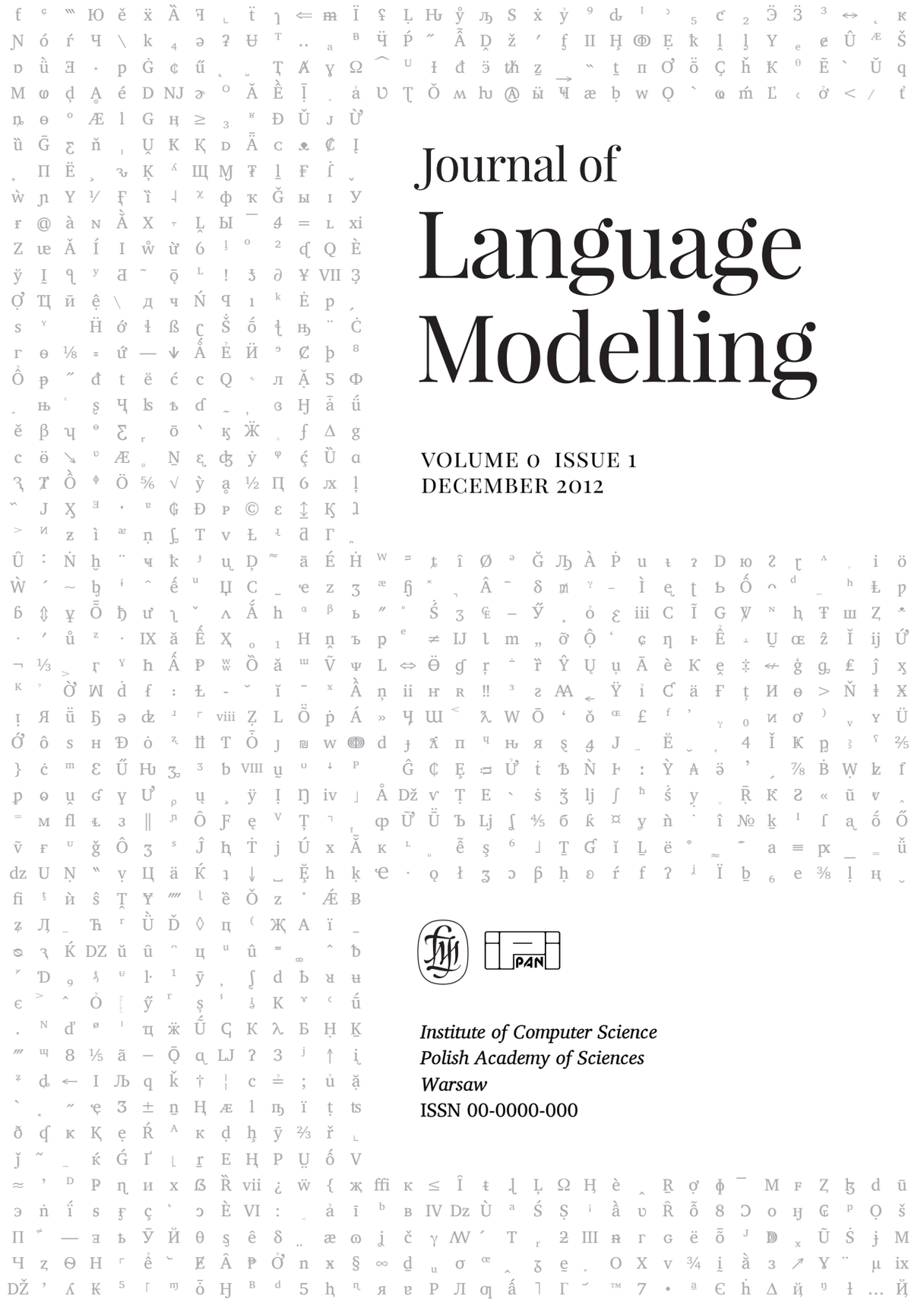
Finally, classification accuracy is measured as the ratio of correctly classified phonological phrase boundaries (tp_{cc}) versus all true positive phonological phrase boundaries (tp):

$$\text{Acc} = \frac{tp_{cc}}{tp}. \tag{4}$$

Classification accuracy was found to equal overall 73.1%.

Varia

Cover



Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science
Polish Academy of Sciences
Warsaw
ISSN 00-0000-000

Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science
Polish Academy of Sciences
Warsaw
ISSN 00-0000-000

Journal of Language Modelling

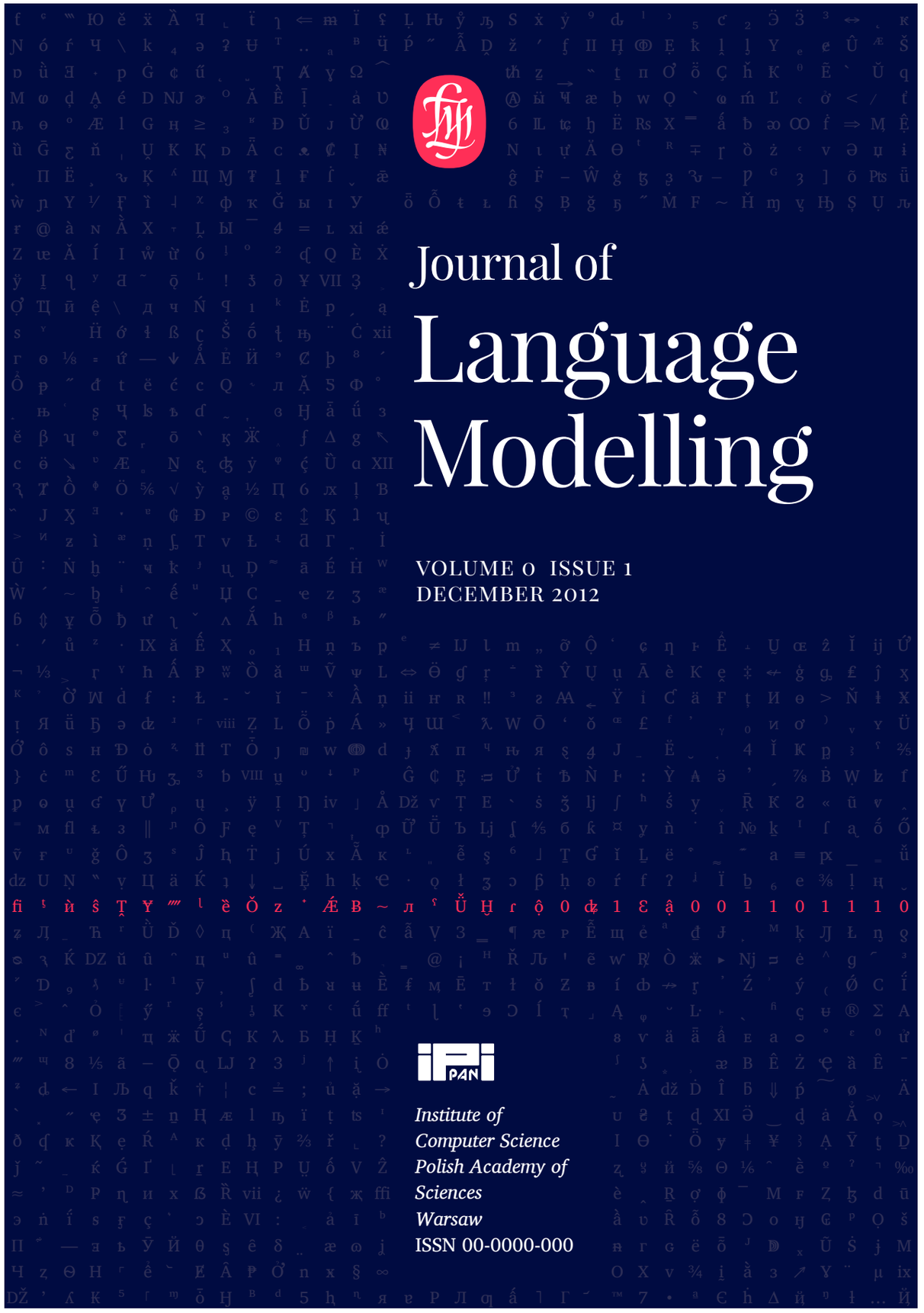
VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science Polish Academy of Sciences Warsaw
ISSN 00-0000-000

Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science Polish Academy of Sciences Warsaw
ISSN 00-0000-000



Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute of Computer Science Polish Academy of Sciences Warsaw
ISSN 00-0000-000

Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Institute
of Computer Science
Polish Academy
of Sciences
Warsaw
ISSN 00-0000-000


Journal of
Language
Modelling
VOLUME 0 ISSUE 1
DECEMBER 2012
Institute of Computer Science
Polish Academy of Sciences
Warsaw

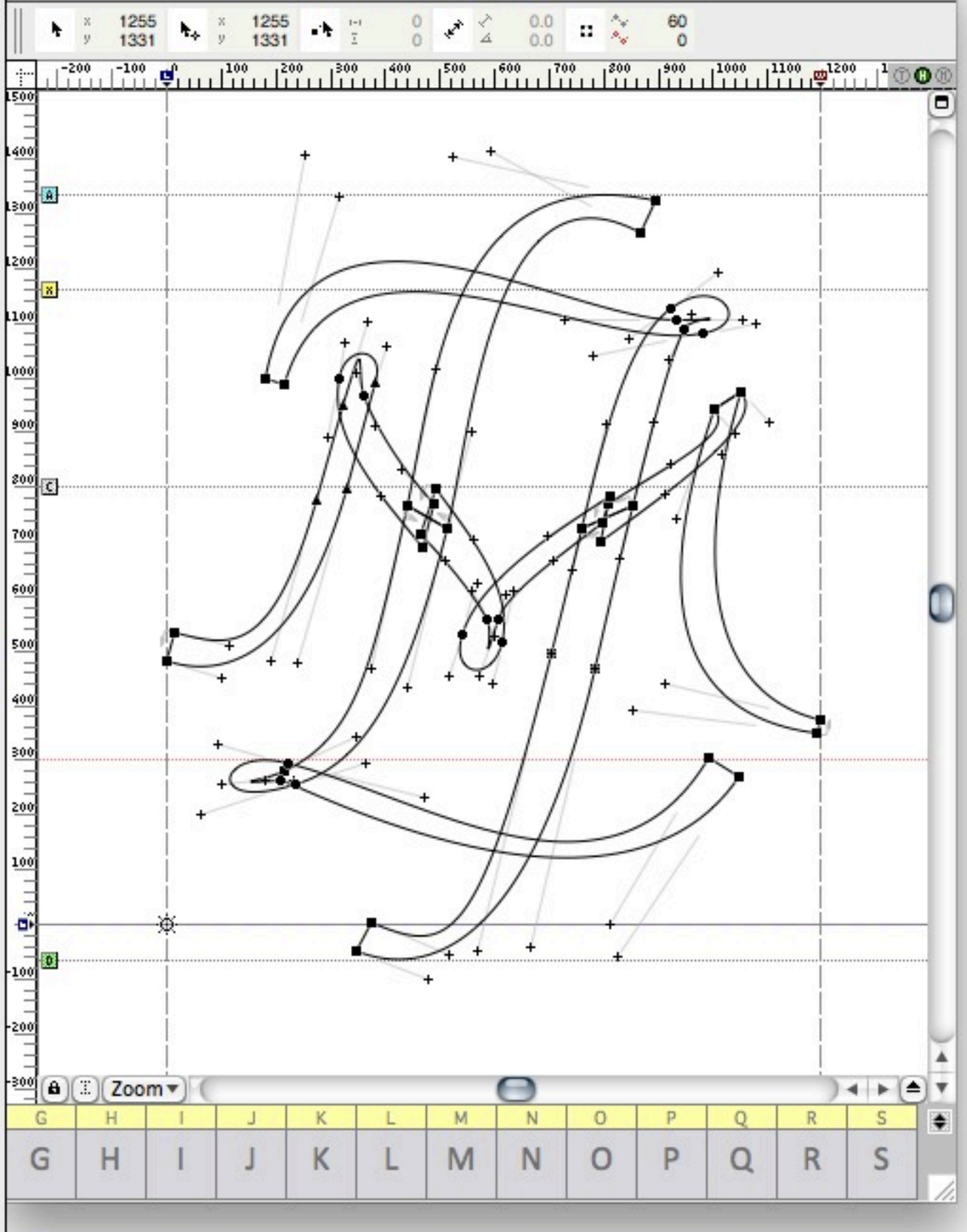
Journal of Language Modelling

VOLUME 0 ISSUE 1

FontLab Studio File Edit View Contour Glyph Tools Window Help

CanvasIterat Top level FMX Class Walk

Glyph – a from JLMLogo3 Regular



Adam

Łukasz Dziedzic

Journal of Language Modelling

VOLUME 0 ISSUE 1
DECEMBER 2012

Editorials

Journal of Language Modelling 1
Adam Przepiórkowski

The Case for the Journal's Use of a CC-BY License 5
Stuart M. Shieber

A Personal Note on Open Access in Linguistics 9
Stefan Müller

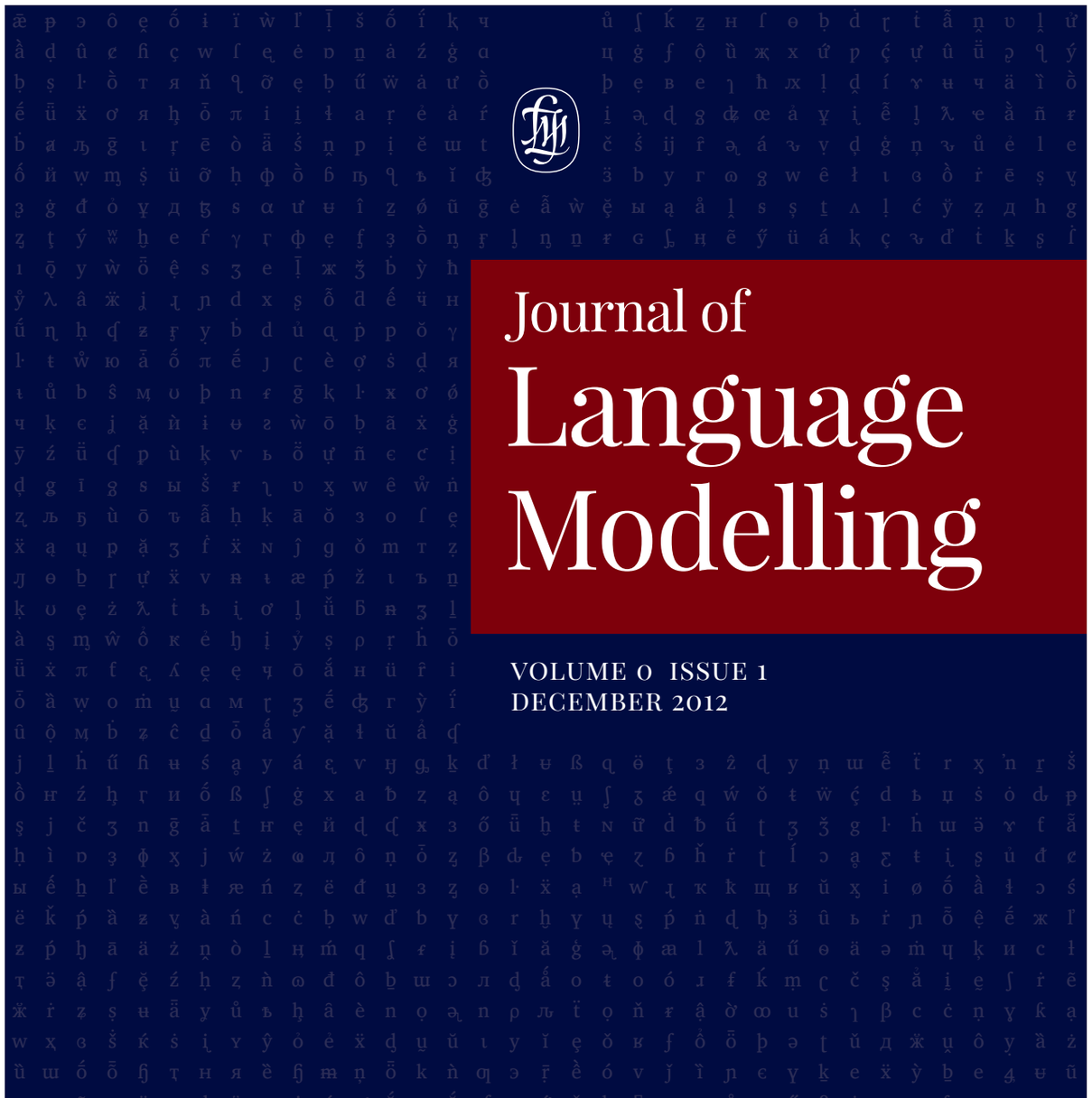
Articles

Slovak Morphosyntactic Tagset 41
Radovan Garabík, Mária Šimková

The Bulgarian National Corpus:
Theory and Practice in Corpus Design 65
Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, Ekaterina Tarpomanova

Derivational and Semantic Relations of Croatian Verbs 111
Krešimir Šojat, Matea Srebačić, Marko Tadić

Exploiting Prosody for Automatic
Syntactic Phrase Boundary Detection in Speech 143
György Szaszák, András Beke



Charis & Charlet

$$nx^{5}\ ba^{ab^{a}}$$

$$nx^{5}\ ba^{ab^{a}}$$

Charis & Charlet

JOHamburgefon Charis SIL

JOHamburgefo Charlet SL S

JOHamburgef Charlet SL XS

Charis & Cambria

JOHamburgefons0123
JOHamburgefons0123

JOHamburgefons0123
JOHamburgefons0123

0000	***	0002	***	***	***	***	***	***	0009	000A	***	***	000D	***	***	0110	0111	0112	0113	0114	0115	0116	0117	0118	0119	011A	011B	011C	011D	011E	011F
																Đ	đ	Ē	ē	Ě	ě	Ė	ė	Ę	ę	Ě	ě	Ĝ	ĝ	Ğ	ğ
***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	0120	0121	0122	0123	0124	0125	0126	0127	0128	0129	012A	012B	012C	012D	012E	012F
																Ġ	ġ	Ģ	ģ	Ĥ	ĥ	Ħ	ħ	Ĩ	ĩ	Ī	ī	Ĭ	ĭ	Į	į
0020	0021	0022	0023	0024	0025	0026	0027	0028	0029	002A	002B	002C	002D	002E	002F	0130	0132	0133	0134	0135	0136	0137	0138	0139	013A	013B	013C	013D	013E	013F	0140
	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/	İ	IJ	ij	Ĵ	ĵ	Ķ	ķ	ĸ	Ĺ	Í	Ļ	ļ	Ľ	ľ	L·	l·
0030	0031	0032	0033	0034	0035	0036	0037	0038	0039	003A	003B	003C	003D	003E	003F	0141	0142	0143	0144	0145	0146	0147	0148	0149	014A	014B	014C	014D	014E	014F	0150
0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?	Ł	ł	Ń	ń	Ņ	ņ	Ň	ň	ʼn	Ŋ	ŋ	Ō	ō	Ŏ	ŏ	Ő
0040	0041	0042	0043	0044	0045	0046	0047	0048	0049	004A	004B	004C	004D	004E	004F	0151	0154	0155	0156	0157	0158	0159	015A	015B	015C	015D	015E	015F	0160	0161	0162
@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	ő	Ŕ	ŕ	Ŗ	ŗ	Ř	ř	Ś	ś	Ŝ	ŝ	Ş	ş	Š	š	Ţ
0050	0051	0052	0053	0054	0055	0056	0057	0058	0059	005A	005B	005C	005D	005E	005F	0163	0164	0165	0166	0167	0168	0169	016A	016B	016C	016D	016E	016F	0170	0171	0172
P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_	ţ	Ť	ť	Ŧ	ŧ	Ũ	ũ	Ū	ū	Ŭ	ŭ	Ů	ů	Ű	ű	Ų
0060	0061	0062	0063	0064	0065	0066	0067	0068	0069	006A	006B	006C	006D	006E	006F	0173	0174	0175	0176	0177	0179	017A	017B	017C	017D	017E	017F	0180	0181	0182	0183
`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	ų	Ŵ	ŵ	Ŷ	ŷ	Ź	ź	Ż	ż	Ž	ž	ſ	ƀ	Ɓ	Ƃ	ƃ
0070	0071	0072	0073	0074	0075	0076	0077	0078	0079	007A	007B	007C	007D	007E	***	0184	0185	0186	0187	0188	0189	018A	018B	018C	018D	018E	018F	0190	0191	0193	0194
p	q	r	s	t	u	v	w	x	y	z	{	|	}	~		ƅ	Ɓ	Ɔ	Ƈ	ƈ	Đ	Ɗ	Ƌ	ƌ	ƍ	Ǝ	Ə	Ɛ	Ƒ	Ɠ	Ɣ
00C4	00C5	00C7	00C9	00D1	00D6	00DC	00E1	00E0	00E2	00E4	00E3	00E5	00E7	00E9	00E8	0195	0196	0197	0198	0199	019A	019B	019C	019D	019E	019F	01A0	01A1	01A2	01A3	01A4
Ä	Å	Ç	É	Ñ	Ö	Ü	á	à	â	ä	ã	å	ç	é	è	ƕ	ƚ	ƛ	Ƙ	ƙ	ƚ	ƛ	Ɯ	Ɲ	ƞ	Ɵ	Ơ	ơ	Ƣ	ƣ	Ƥ
00EA	00EB	00ED	00EC	00EE	00EF	00F1	00F3	00F2	00F4	00F6	00F5	00FA	00F9	00FB	00FC	01A5	01A6	01A7	01A8	01A9	01AA	01AB	01AC	01AD	01AE	01AF	01B0	01B1	01B2	01B3	01B4
ê	ë	í	ì	î	ï	ñ	ó	ò	ô	ö	õ	ú	ù	û	ü	ƥ	ƦR	Ƨ	ƨ	Σ	ƪ	ƫ	Ƭ	ƭ	Ư	ư	Ʊ	Ʋ	Ƴ	ƴ
2020	00B0	00A2	00A3	00A7	2022	00B6	00DF	00AE	00A9	2122	00B4	00A8	2260	00C6	00D8	01B5	01B6	01B7	01B8	01B9	01BA	01BB	01BC	01BD	01BE	01BF	01C0	01C1	01C2	01C3	01C4
†	°	¢	£	§	•	¶	ß	®	©	™	´	¨	≠	Æ	Ø	Ƶ	ƶ	Ʒ	Ƹ	ƹ	ƺ	Ƨ	ƽ	Ƽ	ƽ	ƾ	ƿ	ǀ	ǁ	ǂ	ǃ	DŽ
221E	00B1	2264	2265	00A5	00B5	2202	2211	220F	03C0	222B	00AA	00BA	03A9	00E6	00F8	01C5	01C6	01C7	01C8	01C9	01CA	01CB	01CC	01CD	01CE	01CF	01D0	01D1	01D2	01D3	01D4
∞	±	≤	≥	¥	µ	∂	Σ	Π	π	∫	ª	º	Ω	æ	ø	Dž	dž	LJ	Lj	lj	NJ	Nj	nj	Ǎ	ǎ	Ǐ	ǐ	Ǒ	ǒ	Ǔ	ǔ
00BF	00A1	00AC	221A	0192	2248	2206	00AB	00BB	2026	00A0	00C0	00C3	00D5	0152	0153	01D5	01D6	01D7	01D8	01D9	01DA	01DB	01DC	01DD	01DE	01DF	01E0	01E1	01E2	01E3	01E4
¿	¡	¬	√	ƒ	≈	Δ	«	»	…		À	Ã	Õ	Œ	œ	Ǖ	ǖ	Ǘ	ǘ	Ǚ	ǚ	Ǜ	ǜ	ə	Ǟ	ǟ	Ǡ	ǡ	Æ	æ	Ǥ
2013	2014	201C	201D	2018	2019	00F7	25CA	00FF	0178	2044	20AC	2039	203A	FB01	FB02	01E5	01E6	01E7	01E8	01E9	01EA	01EB	01EC	01ED	01EE	01EF	01F0	01F1	01F2	01F3	01F4
–	—	“	”	‘	’	÷	◊	ÿ	Ÿ	⁄	€	‹	›	fi	fl	ǧ	Ǧ	ǧ	Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ	ǰ	DZ	Dz	dz	Ǵ
2021	00B7	201A	201E	2030	00C2	00CA	00C1	00CB	00C8	00CD	00CE	00CF	00CC	00D3	00D4	01F5	01F6	01F7	01F8	01F9	01FA	01FB	01FC	01FD	01FE	01FF	0200	0201	0202	0203	0204
‡	·	‚	„	‰	Â	Ê	Á	Ë	È	Í	Î	Ï	Ì	Ó	Ô	ǵ	Ƕ	Ƿ	Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ	Ȁ	ȁ	Ȃ	ȃ	Ȅ
F8FF	00D2	00DA	00DB	00D9	0131	02C6	02DC	00AF	02D8	02D9	02DA	00B8	02DD	02DB	02C7	0205	0206	0207	0208	0209	020A	020B	020C	020D	020E	020F	0210	0211	0212	0213	0214
	Ò	Ú	Û	Ù	ı	ˆ	˜	¯	˘	˙	°	¸	˝	˛	ˇ	ȅ	Ȇ	ȇ	Ȉ	ȉ	Ȋ	ȋ	Ȍ	ȍ	Ȏ	ȏ	Ȑ	ȑ	Ȓ	ȓ	Ȕ
00A4	00A6	00AD	00B2	00B3	00B9	00BC	00BD	00BE	00D0	00D7	00DD	00DE	00F0	00FD	00FE	0215	0216	0217	0218	0219	021A	021B	021C	021D	021E	021F	0220	0221	0222	0223	0224
¤	¦	-	²	³	¹	¼	½	¾	Đ	×	Ý	Þ	ð	ý	þ	ȕ	Û	û	Ș	ș	Ț	ț	Ȝ	ȝ	Ȟ	ȟ	Ƞ	ȡ	Ȣ	ȣ	Ȥ
0100	0101	0102	0103	0104	0105	0106	0107	0108	0109	010A	010B	010C	010D	010E	010F	0225	0226	0227	0228	0229	022A	022B	022C	022D	022E	022F	0230	0231	0232	0233	0234
Ā	ā	Ă	ă	Ą	ą	Ć	ć	Ĉ	ĉ	Ċ	ċ	Č	č	Ď	ď	ȥ	Ȧ	ȧ	Ȩ	ȩ	Ȫ	ȫ	Ȭ	ȭ	Ȯ	ȯ	Ȱ	ȱ	Ȳ	ȳ	ȴ

0235	0236	0237	0238	0238	0239	0239	023A	023B	023C	023C	023D	023D	023E	023F	023F	0344	0345	0346	0347	0348	0349	034A	034B	034C	034D	034E	034F	0350	0351	0352	0353
ɳ̊	ɬ̊	ɟ̊	ʠ	ʠ	ɸ	ɸ	Ⱥ	Ȼ	ȼ	ȼ	Ł	Ł	Ⱦ	Ɋ	Ɋ	◌̓	◌ͅ	◌̔	◌͇	◌̈́	◌̋	◌̗	◌̼	◌̽	◌̾	◌̿	◌͆	◌͉	◌͊	◌͋	◌͌	◌͍
0240	0240	0241	0242	0242	0243	0243	0244	0244	0245	0245	0246	0246	0247	0247	0248	0354	0355	0356	0357	0358	0359	035A	035B	035C	035C	035D	035E	035F	0360	0361	0362
ɀ	ɀ	ʔ	ʔ	ʔ	Ƀ	Ƀ	Ʉ	Ʉ	Λ	Λ	Ɇ	Ɇ	ɇ	ɇ	ɟ	◌͔	◌͕	◌͖	◌͗	◌͘	◌͙	◌͚	◌͛	◌͜	◌͝	◌͞	◌͟	◌͠	◌͡	◌͢	◌ͣ	◌ͤ
0248	0249	0249	024A	024A	024B	024B	024C	024C	024D	024D	024E	024E	024F	024F	0250	0363	0364	0365	0366	0367	0368	0369	036A	036B	036C	036D	036E	036F	0374	0375	037A
ɟ	ɟ	ɟ	ɋ	ɋ	ɋ	ɋ	Ʀ	Ʀ	ɾ	ɾ	Ɏ	Ɏ	ɏ	ɏ	ɐ	a	e	i	o	u	c	d	h	m	r	t	v	x	◌ͥ	◌ͦ	◌ͧ
0251	0252	0253	0254	0255	0256	0257	0258	0259	025A	025B	025C	025D	025E	025F	0260	037B	037C	037D	037E	0384	0385	0386	0387	0388	0389	038A	038C	038E	038F	0390	0391
ɑ	ɒ	ɓ	ɔ	ɕ	ɖ	ɗ	ɘ	ə	ɚ	ɛ	ɜ	ɝ	ɞ	ɟ	ɠ	ɔ	ɕ	ɘ	;	΄	◌̈́	Ά	·	Έ	Ή	Ί	Ό	Ύ	Ώ	◌̈́	Α
0261	0262	0263	0264	0265	0266	0267	0268	0269	026A	026B	026C	026D	026E	026F	0270	0392	0393	0394	0395	0396	0397	0398	0399	039A	039B	039C	039D	039E	039F	03A0	03A1
ɡ	ɢ	ɣ	ɤ	ɥ	ɦ	ɧ	ɨ	ɪ	ɪ	ɬ	ɬ	ɭ	ɮ	ɯ	ɰ	Β	Γ	Δ	Ε	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο	Π	Ρ
0271	0272	0273	0274	0275	0276	0277	0278	0279	027A	027B	027C	027D	027E	027F	0280	03A3	03A4	03A5	03A6	03A7	03A8	03AA	03AB	03AC	03AD	03AE	03AF	03B0	03B1	03B2	03B3
ɱ	ɲ	ɳ	ɴ	ɵ	ɶ	ɷ	ɸ	ɹ	ɺ	ɻ	ɽ	ɾ	ɿ	ʀ	ʁ	Σ	Τ	Υ	Φ	Χ	Ψ	Ϊ	Ϋ	ά	έ	ή	ί	ΰ	α	β	γ
0281	0282	0283	0284	0285	0286	0287	0288	0289	028A	028B	028C	028D	028E	028F	0290	03B4	03B5	03B6	03B7	03B8	03B9	03BA	03BB	03BC	03BD	03BE	03BF	03C1	03C2	03C3	03C4
ʁ	ʂ	ʃ	ʄ	ʅ	ʆ	ʇ	ʈ	ʉ	ʊ	ʊ	ʌ	ʍ	ʎ	ʏ	ʐ	δ	ε	ζ	η	θ	ι	κ	λ	μ	ν	ξ	ο	ρ	ς	σ	τ
0291	0292	0293	0294	0295	0296	0297	0298	0299	029A	029B	029C	029D	029E	029F	02A0	03C5	03C6	03C7	03C8	03C9	03CA	03CB	03CC	03CD	03CE	03D0	03D1	03D2	03D3	03D4	03D5
ʑ	ʒ	ʒ	ʔ	ʕ	ʖ	ʗ	ʘ	ʙ	ʚ	ʛ	ʜ	ʝ	ʞ	ʟ	ʠ	υ	φ	χ	ψ	ω	ϊ	ϋ	ό	ύ	ώ	ϐ	ϑ	ϒ	ΰ	ϕ	ϖ
02A1	02A2	02A3	02A4	02A5	02A6	02A7	02A8	02A9	02AA	02AB	02AC	02AD	02AE	02AF	02B0	03D6	03D7	03D8	03D9	03DA	03DB	03DC	03DD	03DE	03DF	03E0	03E1	03E2	03E3	03E4	03E5
ʡ	ʢ	ʣ	ʤ	ʥ	ʦ	ʧ	ʨ	ʩ	ʪ	ʫ	ʬ	ʭ	ʮ	ʯ	ʰ	ϖ	ϗ	Ϙ	ϙ	ς	ς	Ϝ	ϝ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ	Ϥ	ϥ
02B1	02B2	02B3	02B4	02B5	02B6	02B7	02B8	02B9	02BA	02BB	02BC	02BD	02BE	02BF	02C0	03E6	03E7	03E8	03E9	03EA	03EB	03EC	03ED	03EE	03EF	03F0	03F1	03F2	03F3	03F4	03F5
ʱ	ʲ	ʳ	ʴ	ʵ	ʶ	ʷ	ʸ	ʹ	ʺ	ʻ	ʼ	ʽ	ʾ	ʿ	ˀ	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ
02C0	02C1	02C2	02C3	02C4	02C5	02C8	02C8	02C9	02CA	02CA	02CB	02CB	02CC	02CD	02CE	03F6	03F7	03F8	03F9	03FA	03FB	03FC	03FD	03FE	03FF	0400	0401	0402	0403	0404	0405
ˀ	ˁ	˂	˃	˄	˅	ˈ	ˈ	ˉ	ˊ	ˊ	ˋ	ˋ	ˌ	ˍ	ˎ	϶	Ϸ	ϸ	Ϲ	Ϻ	ϻ	ϼ	Ͻ	Ͼ	Ͽ	Ѐ	Ё	Ђ	Ѓ	Є	Ѕ
02CF	02D0	02D1	02D2	02D3	02D4	02D5	02D6	02D7	02DE	02DF	02E0	02E1	02E2	02E3	02E4	0406	0407	0408	0409	040A	040B	040C	040D	040E	040F	0410	0411	0412	0413	0414	0415
ˏ	ː	ˑ	˒	˓	˔	˕	˖	˗	˞	˟	ˠ	ˡ	ˢ	ˣ	ˤ	І	Ї	Ј	Љ	Њ	Ћ	Ќ	Й	Ў	Џ	А	Б	В	Г	Д	Е
02E5	02E6	02E7	02E8	02E9	02EA	02EB	02EC	02EC	02ED	02EE	02EF	02F0	02F1	02F2	02F3	0416	0417	0418	0419	041A	041B	041C	041D	041E	041F	0420	0421	0422	0423	0424	0425
˥	˦	˧	˨	˩	˪	˫	ˬ	˭	ˮ	˯	˰	˱	˲	˳	˴	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х
02F4	02F5	02F6	02F7	02F8	02F9	02FA	02FB	02FC	02FD	02FE	02FF	0300	0301	0302	0303	0426	0427	0428	0429	042A	042B	042C	042D	042E	042F	0430	0431	0432	0433	0434	0435
˵	˶	˷	˷	˸	˹	˺	˻	˼	˽	˾	˿	◌̀	◌́	◌̂	◌̃	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	а	б	в	г	д	е
0304	0305	0306	0307	0308	0309	030A	030B	030C	030D	030E	030F	0310	0311	0312	0313	0436	0437	0438	0439	043A	043B	043C	043D	043E	043F	0440	0441	0442	0443	0444	0445
◌̄	◌̅	◌̆	◌̇	◌̈	◌̉	◌̊	◌̋	◌̌	◌̍	◌̎	◌̏	◌̐	◌̑	◌̒	◌̓	ж	з	и	й	к	л	м	н	о	п	р	с	т	у	ф	х
0314	0315	0316	0317	0318	0319	031A	031B	031C	031D	031E	031F	0320	0321	0322	0323	0446	0447	0448	0449	044A	044B	044C	044D	044E	044F	0450	0451	0452	0453	0454	0455
◌̔	◌̕	◌̖	◌̗	◌̘	◌̙	◌̚	◌̛	◌̜	◌̝	◌̞	◌̟	◌̠	◌̡	◌̢	◌̣	ц	ч	ш	щ	ъ	ы	ь	э	ю	я	ѐ	ё	ђ	ѓ	є	ѕ
0324	0325	0326	0327	0328	0329	032A	032B	032C	032D	032E	032F	0330	0331	0332	0333	0456	0457	0458	0459	045A	045B	045C	045D	045E	045F	0460	0461	0462	0463	0464	0465
◌̤	◌̥	◌̦	◌̧	◌̨	◌̩	◌̪	◌̫	◌̬	◌̭	◌̮	◌̯	◌̰	◌̱	◌̲	◌̳	і	ї	ј	љ	њ	ћ	ќ	й	ў	џ	Ѡ	ѡ	Ѣ	ѣ	Ѥ	ѥ
0334	0335	0336	0337	0338	0339	033A	033B	033C	033D	033E	033F	0340	0341	0342	0343	0466	0467	0468	0469	046A	046B	046C	046D	046E	046F	0470	0471	0472	0473	0474	0475
◌̴	◌̵	◌̶	◌̷	◌̸	◌̹	◌̺	◌̻	◌̼	◌̽	◌̾	◌̿	◌̀	◌́	◌̂	◌̓	Ѧ	ѧ	Ѩ	ѩ	Ѫ	ѫ	Ѭ	ѭ	Ѯ	ѯ	Ѱ	ѱ	Ѳ	ѳ	Ѵ	ѵ

0476	0477	0478	0479	047A	047B	047C	047D	047E	047F	0480	0481	0482	0483	0484	0485
Ѷ	ѷ	Оу	оу	Ѻ	ѻ	Ѽ	ѽ	Ѿ	ѿ	Ҁ	ҁ	≠	◌҃	◌҄	◌҅
0486	0488	0489	048A	048B	048C	048D	048E	048F	0490	0491	0492	0493	0494	0495	0496
◌҆	◌҈	◌҉	Ҋ	ҋ	Ҍ	ҍ	Ҏ	ҏ	Ґ	ґ	Ғ	ғ	Ҕ	ҕ	Җ
0497	0498	0499	049A	049B	049C	049D	049E	049F	04A0	04A1	04A2	04A3	04A4	04A5	04A6
ж	Ӂ	ӂ	Қ	қ	Ҟ	ҟ	Ҡ	ҡ	Ҡ	ҡ	Ң	ң	Ҥ	ҥ	Ҧ
04A7	04A8	04A9	04AA	04AB	04AC	04AD	04AE	04AF	04B0	04B1	04B2	04B3	04B4	04B5	04B6
ҧ	Ѡ	ѡ	Ҫ	ҫ	Ҭ	ҭ	Ү	ү	Ұ	ұ	Ҳ	ҳ	Ҵ	ҵ	Ҷ
04B7	04B8	04B9	04BA	04BB	04BC	04BD	04BE	04BF	04C0	04C1	04C2	04C3	04C4	04C5	04C6
ҷ	Ҹ	ҹ	Һ	һ	Ҽ	ҽ	Ҿ	ҿ	І	Ӂ	ӂ	Ӄ	ӄ	Ӆ	ӆ
04C7	04C8	04C9	04CA	04CB	04CC	04CD	04CE	04CF	04D0	04D1	04D2	04D3	04D4	04D5	04D6
Ӈ	ӈ	Ӊ	ӊ	Ӌ	ӌ	Ӎ	ӎ	Ӏ	Ӑ	ӑ	Ӓ	ӓ	Ӕ	ӕ	Ӗ
04D7	04D8	04D9	04DA	04DB	04DC	04DD	04DE	04DF	04E0	04E1	04E2	04E3	04E4	04E5	04E6
ӗ	Ә	ә	Ӛ	ӛ	Ӝ	ӝ	Ӟ	ӟ	Ҙ	ҙ	Ӣ	ӣ	Ӥ	ӥ	Ӧ
04E7	04E8	04E9	04EA	04EB	04EC	04ED	04EE	04EF	04F0	04F1	04F2	04F3	04F4	04F5	04F6
ӧ	Ө	ө	Ӫ	ӫ	Ӭ	ӭ	Ӯ	ӯ	Ӱ	ӱ	Ӳ	ӳ	Ӵ	ӵ	Ҷ
04F6	04F7	04F7	04F8	04F9	04FA	04FB	04FC	04FC	04FD	04FD	04FE	04FE	04FF	04FF	0500
ҷ	Ҹ	ҹ	Ӹ	ӹ	Ҩ	ҩ	Ҳ	Ҳ	ҳ	ҳ	Ҵ	Ҵ	ҵ	ҵ	Ԁ
0501	0502	0503	0504	0505	0506	0507	0508	0509	050A	050B	050C	050D	050E	050F	0510
ԁ	Ԃ	ԃ	Ԅ	ԅ	Ԇ	ԇ	Ԉ	ԉ	Ԋ	ԋ	Ԍ	ԍ	Ԏ	ԏ	Ԑ
0510	0511	0511	0512	0512	0513	0513	0514	0515	0516	0517	0518	0519	051A	051B	051C
Ԑ	ԑ	ԑ	Ԓ	Ԓ	Ԓ	Ԓ	Ԕ	Ԕ	Ԗ	Ԗ	Ԙ	Ԙ	Q	q	W
051D	051E	051F	0520	0521	0522	0523	0524	0525	0526	0527	0E3F	1D00	1D01	1D02	1D03
w	Ԟ	ԟ	Ԡ	ԡ	Ԣ	ԣ	Ԥ	ԥ	Ԧ	ԧ	฿	ᴀ	ᴁ	ᴂ	ʙ
1D04	1D05	1D06	1D07	1D08	1D09	1D0A	1D0B	1D0C	1D0D	1D0E	1D0F	1D10	1D11	1D12	1D13
ᴄ	ᴅ	ᴆ	ᴇ	ᴈ	ᴉ	ᴊ	ᴋ	ᴌ	ᴍ	ᴎ	ᴏ	ᴐ	ᴑ	ᴒ	ᴓ
1D14	1D15	1D16	1D17	1D18	1D19	1D1A	1D1B	1D1C	1D1D	1D1E	1D1F	1D20	1D21	1D22	1D23
ᴔ	ᴕ	ᴖ	ᴗ	ᴘ	ᴙ	ᴚ	ᴛ	ᴜ	ᴝ	ᴞ	ᴟ	ᴠ	ᴡ	ᴢ	ᴣ
1D24	1D25	1D26	1D27	1D28	1D29	1D2A	1D2B	1D2C	1D2D	1D2E	1D2F	1D30	1D31	1D32	1D33
ᴤ	ᴥ	ᴦ	ᴧ	ᴨ	ᴩ	ᴪ	ᴫ	ᴬ	ᴭ	ᴮ	ᴯ	ᴰ	ᴱ	ᴲ	ᴳ
1D34	1D35	1D36	1D37	1D38	1D39	1D3A	1D3B	1D3C	1D3D	1D3E	1D3F	1D40	1D41	1D42	1D43
ᴴ	ᴵ	ᴶ	ᴷ	ᴸ	ᴹ	ᴺ	ᴻ	ᴼ	ᴽ	ᴾ	ᴿ	ᵀ	ᵁ	ᵂ	ᵃ
1D44	1D45	1D46	1D47	1D48	1D49	1D4A	1D4B	1D4C	1D4D	1D4E	1D4F	1D50	1D50	1D51	1D51
ᵄ	ᵅ	ᵆ	ᵇ	ᵈ	ᵉ	ᵊ	ᵋ	ᵌ	ᵍ	ᵎ	ᵏ	m	m	ŋ	ŋ
1D52	1D53	1D54	1D55	1D56	1D57	1D58	1D59	1D5A	1D5B	1D5C	1D5D	1D5E	1D5F	1D60	1D61
ᵒ	ᵓ	ᵔ	ᵕ	ᵖ	ᵗ	ᵘ	ᵙ	ᵚ	ᵛ	ᵜ	ᵝ	ᵞ	ᵟ	ᵠ	ᵡ

1D62	1D63	1D64	1D65	1D66	1D67	1D68	1D69	1D6A	1D6B	1D6C	1D6C	1D6D	1D6D	1D6E	1D6E
ᵢ	ᵣ	ᵤ	ᵥ	ᵦ	ᵧ	ᵨ	ᵩ	ᵪ	ᵫ	ƀ	ƀ	đ	đ	ɸ	ɸ
1D6F	1D6F	1D70	1D70	1D71	1D71	1D72	1D72	1D73	1D73	1D74	1D74	1D75	1D75	1D76	1D76
ᵯ	ᵯ	ᵰ	ᵰ	ᵱ	ᵱ	ᵲ	ᵲ	ᵳ	ᵳ	ᵴ	ᵴ	ᵵ	ᵵ	ᵶ	ᵶ
1D77	1D78	1D79	1D7A	1D7B	1D7B	1D7C	1D7C	1D7D	1D7D	1D7E	1D7E	1D7F	1D7F	1D80	1D80
ᵷ	ᵸ	ᵹ	ᵺ	ᵻ	ᵻ	ᵼ	ᵼ	ᵽ	ᵽ	ᵾ	ᵾ	ᵿ	ᵿ	ᶀ	ᶀ
1D81	1D81	1D82	1D82	1D83	1D83	1D84	1D84	1D85	1D85	1D86	1D86	1D87	1D87	1D88	1D88
ᶁ	ᶁ	ᶂ	ᶂ	ᶃ	ᶃ	ᶄ	ᶄ	ᶅ	ᶅ	ᶆ	ᶆ	ᶇ	ᶇ	ᶈ	ᶈ
1D89	1D89	1D8A	1D8A	1D8B	1D8B	1D8C	1D8C	1D8D	1D8D	1D8E	1D8E	1D8F	1D8F	1D90	1D90
ᶉ	ᶉ	ᶊ	ᶊ	ᶋ	ᶋ	ᶌ	ᶌ	ᶍ	ᶍ	ᶎ	ᶎ	ᶏ	ᶏ	ᶐ	ᶐ
1D91	1D91	1D92	1D92	1D93	1D93	1D94	1D94	1D95	1D95	1D96	1D96	1D97	1D97	1D98	1D98
ᶑ	ᶑ	ᶒ	ᶒ	ᶓ	ᶓ	ᶔ	ᶔ	ᶕ	ᶕ	ᶖ	ᶖ	ᶗ	ᶗ	ᶘ	ᶘ
1D99	1D99	1D9A	1D9A	1D9B	1D9B	1D9C	1D9C	1D9D	1D9D	1D9E	1D9E	1D9F	1D9F	1DA0	1DA0
ᶙ	ᶙ	ᶚ	ᶚ	ᶛ	ᶛ	ᶜ	ᶜ	ᶝ	ᶝ	ᶞ	ᶞ	ᶟ	ᶟ	ᶠ	ᶠ
1DA1	1DA1	1DA2	1DA2	1DA3	1DA3	1DA4	1DA4	1DA5	1DA5	1DA6	1DA6	1DA7	1DA7	1DA8	1DA8
ᶡ	ᶡ	ᶢ	ᶢ	ᶣ	ᶣ	ᶤ	ᶤ	ᶥ	ᶥ	ᶦ	ᶦ	ᶧ	ᶧ	ᶨ	ᶨ
1DA9	1DA9	1DAA	1DAA	1DAB	1DAB	1DAC	1DAC	1DAD	1DAD	1DAE	1DAE	1DAF	1DAF	1DB0	1DB0
ᶩ	ᶩ	ᶪ	ᶪ	ᶫ	ᶫ	ᶬ	ᶬ	ᶭ	ᶭ	ᶮ	ᶮ	ᶯ	ᶯ	ᶰ	ᶰ
1DB1	1DB1	1DB2	1DB2	1DB3	1DB3	1DB4	1DB4	1DB5	1DB5	1DB6	1DB6	1DB7	1DB7	1DB8	1DB8
ᶱ	ᶱ	ᶲ	ᶲ	ᶳ	ᶳ	ᶴ	ᶴ	ᶵ	ᶵ	ᶶ	ᶶ	ᶷ	ᶷ	ᶸ	ᶸ
1DB9	1DB9	1DBA	1DBA	1DBB	1DBB	1DBC	1DBC	1DBD	1DBD	1DBE	1DBE	1DBF	1DBF	1DC0	1DC1
ᶹ	ᶹ	ᶺ	ᶺ	ᶻ	ᶻ	ᶼ	ᶼ	ᶽ	ᶽ	ᶾ	ᶾ	ᶿ	ᶿ	˟	˟
1DC2	1DC2	1DC3	1DC4	1DC4	1DC5	1DC5	1DC6	1DC6	1DC7	1DC7	1DC8	1DC8	1DC9	1DC9	1DCA
◌᷂	◌᷂	◌᷃	◌᷄	◌᷄	◌᷅	◌᷅	◌᷆	◌᷆	◌᷇	◌᷇	◌᷈	◌᷈	◌᷉	◌᷉	r
1DCA	1DCB	1DCC	1DCD	1DFD	1DFE	1DFF	1E00	1E01	1E02	1E03	1E04	1E05	1E06	1E07	1E08
r	◌᷋	◌᷌	◌᷍	◌᷽	◌᷾	◌᷿	Ḁ	ḁ	Ḃ	ḃ	Ḅ	ḅ	Ḇ	ḇ	Ḉ
1E09	1E0A	1E0B	1E0C	1E0D	1E0E	1E0F	1E10	1E11	1E12	1E13	1E14	1E15	1E16	1E17	1E18
ḉ	Ḋ	ḋ	Ḍ	ḍ	Ḏ	ḏ	Ḑ	ḑ	Ḓ	ḓ	Ḕ	ḕ	Ḗ	ḗ	Ḙ
1E19	1E1A	1E1B	1E1C	1E1D	1E1E	1E1F	1E20	1E21	1E22	1E23	1E24	1E25	1E26	1E27	1E28
ḙ	Ḛ	ḛ	Ḝ	ḝ	Ḟ	ḟ	Ḡ	ḡ	Ḣ	ḣ	Ḥ	ḥ	Ḧ	ḧ	Ḩ
1E29	1E2A	1E2B	1E2C	1E2D	1E2E	1E2F	1E30	1E31	1E32	1E33	1E34	1E35	1E36	1E37	1E38
ḩ	Ḫ	ḫ	Ḭ	ḭ	Ḯ	ḯ	Ḱ	ḱ	Ḳ	ḳ	Ḵ	ḵ	Ḷ	ḷ	Ḹ
1E39	1E3A	1E3B	1E3C	1E3D	1E3E	1E3F	1E40	1E41	1E42	1E43	1E44	1E45	1E46	1E47	1E48
ḹ	Ḻ	ḻ	Ḽ	ḽ	Ḿ	ḿ	Ṁ	ṁ	Ṃ	ṃ	Ṅ	ṅ	Ṇ	ṇ	Ṉ
1E49	1E4A	1E4B	1E4C	1E4D	1E4E	1E4F	1E50	1E51	1E52	1E53	1E54	1E55	1E56	1E57	1E58
ṉ	Ṋ	ṋ	Ṍ	ṍ	Ṏ	ṏ	Ṑ	ṑ	Ṓ	ṓ	Ṕ	ṕ	Ṗ	ṗ	Ṙ

1E59	1E5A	1E5B	1E5C	1E5D	1E5E	1E5F	1E60	1E61	1E62	1E63	1E64	1E65	1E66	1E67	1E68	1F87	1F88	1F89	1F8A	1F8B	1F8C	1F8D	1F8E	1F8F	1F90	1F91	1F92	1F93	1F94	1F95	1F96
ṙ	Ṛ	ṛ	Ṝ	ṝ	Ṟ	ṟ	Ṡ	ṡ	Ṣ	ṣ	Ṥ	ṥ	Ṧ	ṧ	Ṩ	ᾇ	ᾈ	ᾉ	ᾊ	ᾋ	ᾌ	ᾍ	ᾎ	ᾏ	ᾗ	ᾘ	ᾙ	ᾚ	ᾛ	ᾜ	ᾝ	ᾞ
1E69	1E6A	1E6B	1E6C	1E6D	1E6E	1E6F	1E70	1E71	1E72	1E73	1E74	1E75	1E76	1E77	1E78	1F97	1F98	1F99	1F9A	1F9B	1F9C	1F9D	1F9E	1F9F	1FA0	1FA1	1FA2	1FA3	1FA4	1FA5	1FA6
ṩ	Ṫ	ṫ	Ṭ	ṭ	Ṯ	ṯ	Ṱ	ṱ	Ṳ	ṳ	Ṵ	ṵ	Ṷ	ṷ	Ṹ	ᾟ	ᾨ	ᾩ	ᾪ	ᾫ	ᾬ	ᾭ	ᾮ	ᾯ	ᾧ	ᾠ	ᾡ	ᾢ	ᾣ	ᾤ	ᾥ	ᾦ
1E79	1E7A	1E7B	1E7C	1E7D	1E7E	1E7F	1E80	1E81	1E82	1E83	1E84	1E85	1E86	1E87	1E88	1FA7	1FA8	1FA9	1FAA	1FAB	1FAC	1FAD	1FAE	1FAF	1FB0	1FB1	1FB2	1FB3	1FB4	1FB6	1FB7
ṹ	Ṻ	ṻ	Ṽ	ṽ	Ṿ	ṿ	Ẁ	ẁ	Ẃ	ẃ	Ẅ	ẅ	Ẇ	ẇ	Ẉ	ᾧ	ᾨ	ᾩ	ᾪ	ᾫ	ᾬ	ᾭ	ᾮ	ᾯ	ᾰ	ᾱ	ᾲ	ᾳ	ᾴ	ᾶ	ᾷ
1E89	1E8A	1E8B	1E8C	1E8D	1E8E	1E8F	1E90	1E91	1E92	1E93	1E94	1E95	1E96	1E97	1E98	1FB8	1FB9	1FBA	1FBB	1FBC	1FBD	1FBE	1FBF	1FC0	1FC1	1FC2	1FC3	1FC4	1FC6	1FC7	1FC8
ẉ	Ẋ	ẋ	Ẍ	ẍ	Ẏ	ẏ	Ẑ	ẑ	Ẓ	ẓ	Ẕ	ẕ	ẖ	ẗ	ẘ	Ᾰ	Ᾱ	Ὰ	Ά	Ᾱ	᾽	ι	᾿	῀	῁	ῂ	ῃ	ῄ	ῆ	ῇ	Ὲ
1E99	1E9A	1E9B	1E9C	1E9D	1E9E	1E9F	1EA0	1EA1	1EA2	1EA3	1EA4	1EA5	1EA6	1EA7	1EA8	1FC9	1FCA	1FCB	1FCC	1FCD	1FCE	1FCF	1FD0	1FD1	1FD2	1FD3	1FD6	1FD7	1FD8	1FD9	1FDA
ẙ	ẚ	ẛ	ẜ	ẝ	ẞ	ẟ	Ạ	ạ	Ả	ả	Ấ	ấ	Ầ	ầ	Ẩ	Έ	Ὴ	Ή	Ῐ	΅	΄	῍	῎	῏	ῐ	ῑ	ῒ	ΐ	ῖ	ῗ	Ὶ
1EA9	1EAA	1EAB	1EAC	1EAD	1EAE	1EAF	1EB0	1EB1	1EB2	1EB3	1EB4	1EB5	1EB6	1EB7	1EB8	1FDB	1FDD	1FDE	1FDF	1FE0	1FE1	1FE2	1FE3	1FE4	1FE5	1FE6	1FE7	1FE8	1FE9	1FEA	1FEB
ẩ	Ẫ	ẫ	Ậ	ậ	Ắ	ắ	Ằ	ằ	Ẳ	ẳ	Ẵ	ẵ	Ặ	ặ	Ẹ	Ί	῝	῞	῟	Ῠ	Ῡ	Ϋ̀	Ϋ́	ῤ	ῥ	ῦ	ῧ	Ὺ	Ύ	Ῥ	Ὼ
1EB9	1EBA	1EBB	1EBC	1EBD	1EBE	1EBF	1EC0	1EC1	1EC2	1EC3	1EC4	1EC5	1EC6	1EC7	1EC8	1FEC	1FED	1FEE	1FEF	1FF2	1FF3	1FF4	1FF6	1FF7	1FF8	1FF9	1FFA	1FFB	1FFC	1FFD	1FFE
ẹ	Ẻ	ẻ	Ẽ	ẽ	Ế	ế	Ề	ề	Ể	ể	Ễ	ễ	Ệ	ệ	Ỉ	ῼ	´	῾	`	ῲ	ῳ	ῴ	ῶ	ῷ	Ὸ	Ό	Ὼ	Ώ	Ω	´	῾
1EC9	1ECA	1ECB	1ECC	1ECD	1ECE	1ECF	1ED0	1ED1	1ED2	1ED3	1ED4	1ED5	1ED6	1ED7	1ED8	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	200A	200B	200C	200D	200E	200F
ỉ	Ị	ị	Ọ	ọ	Ỏ	ỏ	Ố	ố	Ồ	ồ	Ổ	ổ	Ỗ	ỗ	Ộ																
1ED9	1EDA	1EDB	1EDC	1EDD	1EDE	1EDF	1EE0	1EE1	1EE2	1EE3	1EE4	1EE5	1EE6	1EE7	1EE8	2010	2011	2012	2015	2016	2017	201B	201F	2023	2024	2025	2027	2028	2029	202A	202B
ộ	Ớ	ớ	Ờ	ờ	Ở	ở	Ỡ	ỡ	Ợ	ợ	Ụ	ụ	Ủ	ủ	Ứ	‐	‑	‒	—	‖	‗	‛	‟	▶	.	..	·				
1EE9	1EEA	1EEB	1EEC	1EED	1EEE	1EEF	1EF0	1EF1	1EF2	1EF3	1EF4	1EF5	1EF6	1EF7	1EF8	202C	202D	202E	202F	2032	2033	2034	2035	2036	2037	2038	203C	203D	203E	203F	2040
ứ	Ừ	ừ	Ử	ử	Ữ	ữ	Ự	ự	Ỳ	ỳ	Ỵ	ỵ	Ỷ	ỷ	Ỹ					′	″	‴	‵	‶	‷	‸	‼	‽	‾	‿	⁀
1EF9	1EFA	1EFB	1EFC	1EFD	1EFE	1EFF	1F00	1F01	1F02	1F03	1F04	1F05	1F06	1F07	1F08	2053	2057	205E	205F	2060	2061	2062	2063	206A	206B	206C	206D	206E	206F	2070	2071
ỹ	Ⱡ	ⱡ	Ɑ	Ɑ	Ɣ	ɣ	ά	ὰ	ᾀ	ᾁ	ᾂ	ᾃ	ᾄ	ᾅ	ᾆ	~	⁗	⁞												0	i
1F09	1F0A	1F0B	1F0C	1F0D	1F0E	1F0F	1F10	1F11	1F12	1F13	1F14	1F15	1F18	1F19	1F1A	2074	2075	2076	2077	2078	2079	207A	207B	207C	207D	207E	207F	2080	2081	2082	2083
Ἁ	Ἂ	Ἃ	Ἄ	Ἅ	Ἆ	Ἇ	έ	ὲ	ἐ	ἑ	ἒ	ἓ	Ἐ	Ἑ	Ἒ	4	5	6	7	8	9	⁺	⁻	⁼	⁽	⁾	ⁿ	₀	₁	₂	₃
1F1B	1F1C	1F1D	1F20	1F21	1F22	1F23	1F24	1F25	1F26	1F27	1F28	1F29	1F2A	1F2B	1F2C	2084	2085	2086	2087	2088	2089	208A	208B	208C	208D	208E	2090	2091	2092	2093	2094
Ἃ	Ἓ	Ἕ	ή	ὴ	ἠ	ἡ	ἢ	ἣ	ἤ	ἥ	ἦ	ἧ	Ἠ	Ἡ	Ἢ	₄	₅	₆	₇	₈	₉	₊	₋	₌	₍	₎	ₐ	ₑ	ₒ	ₓ	ₔ
1F2D	1F2E	1F2F	1F30	1F31	1F32	1F33	1F34	1F35	1F36	1F37	1F38	1F39	1F3A	1F3B	1F3C	20A0	20A1	20A2	20A3	20A4	20A5	20A6	20A7	20A8	20A9	20AA	20AB	20AD	20AE	20AF	20B0
Ἅ	Ἕ	Ἇ	ί	ὶ	ἰ	ἱ	ἲ	ἳ	ἴ	ἵ	ἶ	ἷ	Ἰ	Ἱ	Ἲ	₠	₡	₢	₣	₤	₥	₦	Pts	Rs	₩	₪	₫	₭	₮	₯	₰
1F3D	1F3E	1F3F	1F40	1F41	1F42	1F43	1F44	1F45	1F48	1F49	1F4A	1F4B	1F4C	1F4D	1F50	20B1	20B2	20B3	20B4	20B5	20B9	20DD	20E5	20EC	20ED	20EE	20EF	2105	2113	2116	2117
Ἑ	Ἵ	Ἷ	ό	ὸ	ὀ	ὁ	ὂ	ὅ	Ὸ	Ό	Ὺ	Ύ	Ὼ	Ώ	ὐ	₱	₲	₳	₴	₵	₹	◯	⃥	⃬	⃭	⃮	⃯	℅	ℓ	№	℗
1F51	1F52	1F53	1F54	1F55	1F56	1F57	1F59	1F5B	1F5D	1F5F	1F60	1F61	1F62	1F63	1F64	2118	2119	211A	211F	2123	2126	212E	2132	2139	214D	214E	2153	2154	2155	2156	2157
ὑ	ὒ	ὓ	ὔ	ὕ	ὖ	ὗ	Ὑ	Ὓ	Ὕ	Ὗ	ὠ	ὡ	ὢ	ὣ	ὤ	℘	ℙ	ℚ	℟	℣	Ω	℮	Ⅎ	ⅈ	⅍	ⅎ	⅓	⅔	⅕	⅖	⅗
1F65	1F66	1F67	1F68	1F69	1F6A	1F6B	1F6C	1F6D	1F6E	1F6F	1F70	1F71	1F72	1F73	1F74	2158	2159	215A	215B	215C	215D	215E	215F	2160	2161	2162	2163	2164	2165	2166	2167
ὥ	ὦ	ὧ	Ὠ	Ὡ	Ὢ	Ὣ	Ὤ	Ὥ	Ὦ	Ὧ	ὰ	ά	ὲ	έ	ὴ	⅘	⅙	⅚	⅛	⅜	⅝	⅞	⅟	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ
1F75	1F76	1F77	1F78	1F79	1F7A	1F7B	1F7C	1F7D	1F80	1F81	1F82	1F83	1F84	1F85	1F86	2168	2169	216A	216B	216C	216D	216E	216F	2170	2171	2172	2173	2174	2175	2176	2177
ή	ὶ	ί	ὸ	ό	ὺ	ύ	ὼ	ώ	ᾀ	ᾁ	ᾂ	ᾃ	ᾄ	ᾅ	ᾆ	Ⅸ	Ⅹ	Ⅺ	Ⅻ	Ⅼ	Ⅽ	Ⅾ	Ⅿ	ⅰ	ⅱ	ⅲ	ⅳ	ⅴ	ⅵ	ⅶ	ⅷ

2178	2179	217A	217B	217C	217D	217E	217F	2180	2181	2182	2183	2184	2185	2186	2187
ⅸ	ⅹ	ⅺ	ⅻ	ↄ	ↄ	ↄ	ↄ	ↀ	ↁ	ↂ	Ↄ	ↄ	ↅ	ↆ	ↇ
2188	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	219A	219B	219C	219D	219E
ↈ	←	↑	→	↓	↔	↕	↖	↗	↘	↙	↚	↛	↜	↝	↞
219F	21A0	21A1	21A2	21A3	21A4	21A5	21A6	21A7	21A8	21A9	21AA	21AB	21AC	21AD	21AE
↟	↠	↡	↢	↣	↤	↥	↦	↧	↨	↩	↪	↫	↬	↭	↮
21AF	21B0	21B1	21B2	21B3	21B4	21B5	21B6	21B7	21B8	21B9	21BA	21BB	21BC	21BD	21BE
↯	↰	↱	↲	↳	↴	↵	↶	↷	↸	↹	↺	↻	↼	↽	↾
21BF	21C0	21C1	21C2	21C3	21C4	21C5	21C6	21C7	21C8	21C9	21CA	21CB	21CC	21CD	21CE
↿	⇀	⇁	⇂	⇃	⇄	⇅	⇆	⇇	⇈	⇉	⇊	⇋	⇌	⇍	⇎
21CF	21D0	21D1	21D2	21D3	21D4	21D5	21D6	21D7	21D8	21D9	21DA	21DB	21DC	21DD	21DE
⇏	⇐	⇑	⇒	⇓	⇔	⇕	⇖	⇗	⇘	⇙	⇚	⇛	⇜	⇝	⇞
21DF	21E0	21E1	21E2	21E3	21E4	21E5	21E6	21E7	21E8	21E9	21EA	21EB	21EC	21ED	21EE
⇟	⇠	⇡	⇢	⇣	⇤	⇥	⇦	⇧	⇨	⇩	⇪	⇫	⇬	⇭	⇮
21EF	21F0	21F1	21F2	21F3	21F4	21F5	21F6	21F7	21F8	21F9	21FA	21FB	21FC	21FD	21FE
⇯	⇰	⇱	⇲	⇳	⇴	⇵	⇶	⇷	⇸	⇹	⇺	⇻	⇼	⇽	⇾
21FF	2200	2201	2203	2204	2205	2208	2209	220A	220B	220C	220D	220E	2210	2212	2213
⇿	∀	∁	∃	∄	∅	∈	∉	∊	∋	∌	∍	∎	∐	−	∓
2214	2215	2216	2217	2218	2219	221B	221C	221D	221F	2220	2221	2222	2223	2224	2225
∔	∕	∖	∗	∘	∙	∛	∜	∝	∟	∠	∡	∢	∣	∤	∥
2226	2227	2228	2229	222A	222C	222D	222E	222F	2230	2231	2232	2233	2234	2235	2236
∦	∧	∨	∩	∪	∬	∭	∮	∯	∰	∱	∲	∳	∴	∵	∶
2237	2238	2239	223A	223B	223C	223D	223E	223F	2240	2241	2242	2243	2244	2245	2246
∷	∸	∹	∺	∻	∼	∽	∾	∿	≀	≁	≂	≃	≄	≅	≆
2247	2249	224A	224B	224C	224D	224E	224F	2250	2251	2252	2253	2254	2255	2256	2257
≇	≉	≊	≋	≌	≍	≎	≏	≐	≑	≒	≓	≔	≕	≖	≗
2258	2259	225A	225B	225C	225D	225E	225F	2261	2262	2263	2266	2267	2268	2269	226A
≘	≙	≚	≛	≜	≝	≞	≟	≡	≢	≣	≦	≧	≨	≩	≪
226B	226C	226D	226E	226F	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	227A
≫	≬	≭	≮	≯	≰	≱	≲	≳	≴	≵	≶	≷	≸	≹	≺
227B	227C	227D	227E	227F	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	228A
≻	≼	≽	≾	≿	⊀	⊁	⊂	⊃	⊄	⊅	⊆	⊇	⊈	⊉	⊊
228B	228C	228D	228E	228F	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	229A
⊋	⊌	⊍	⊎	⊏	⊐	⊑	⊒	⊓	⊔	⊕	⊖	⊗	⊘	⊙	⊚
229B	229C	229D	229E	229F	22A0	22A1	22A2	22A3	22A6	22A7	22A8	22A9	22AA	22AB	22AC
⊛	⊜	⊝	⊞	⊟	⊠	⊡	⊢	⊣	⊦	⊧	⊨	⊩	⊪	⊫	⊬

22AD	22AE	22AF	22B0	22B1	22B2	22B3	22B4	22B5	22B6	22B7	22B8	22B9	22BA	22BB	22BC
⊭	⊮	⊯	⊰	⊱	⊲	⊳	⊴	⊵	⊶	⊷	⊸	⊹	⊺	⊻	⊼
22BD	22BE	22BF	22C0	22C1	22C2	22C3	22C4	22C5	22C6	22C7	22C8	22C9	22CA	22CB	22CC
⊽	⊾	⊿	⋀	⋁	⋂	⋃	⋄	⋅	⋆	⋇	⋈	⋉	⋊	⋋	⋌
22CD	22CE	22CF	22D0	22D1	22D2	22D3	22D4	22D5	22D6	22D7	22D8	22D9	22DA	22DB	22DC
⋍	⋎	⋏	⋐	⋑	⋒	⋓	⋔	⋕	⋖	⋗	⋘	⋙	⋚	⋛	⋜
22DD	22DE	22DF	22E0	22E1	22E2	22E3	22E4	22E5	22E6	22E7	22E8	22E9	22EA	22EB	22EC
⋝	⋞	⋟	⋠	⋡	⋢	⋣	⋤	⋥	⋦	⋧	⋨	⋩	⋪	⋫	⋬
22ED	22EE	22EF	22F0	22F1	22F2	22F3	22F4	22F5	22F6	22F7	22F8	22F9	22FA	22FB	22FC
⋭	⋮	⋯	⋰	⋱	⋲	⋳	⋴	⋵	⋶	⋷	⋸	⋹	⋺	⋻	⋼
22FD	22FE	22FF	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	230A	230B	230C
⋽	⋾	⋿	⌀	⌁	⌂	⌃	⌄	⌅	⌆	⌇	⌈	⌉	⌊	⌋	⌌
230D	230E	230F	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	231A	231B	231C
⌍	⌎	⌏	⌐	⌑	⌒	⌓	⌔	⌕	⌖	⌗	⌘	⌙	⌚	⌛	⌜
231D	231E	231F	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	232A	2330	2331
⌝	⌞	⌟	⌠	⌡	⌢	⌣	⌤	⌥	⌦	⌧	⌨	〈	〉	⌫	⌬
2332	2333	2334	2335	2336	2337	2338	2339	233A	233B	233C	233D	233E	233F	2340	2341
⌭	⌮	⌯	⌰	⌱	⌲	⌳	⌴	⌵	⌶	⌷	⌸	⌹	⌺	⌻	⌼
2342	2343	2344	2345	2346	2347	2348	2349	234A	234B	234C	234D	234E	234F	2350	2351
⍂	⍃	⍄	⍅	⍆	⍇	⍈	⍉	⍊	⍋	⍌	⍍	⍎	⍏	⍐	⍑
2352	2353	2354	2355	2356	2357	2358	2359	235A	235B	235C	235D	235E	235F	2360	2361
⍒	⍓	⍔	⍕	⍖	⍗	⍘	⍙	⍚	⍛	⍜	⍝	⍞	⍟	⍠	⍡
2362	2363	2364	2365	2366	2367	2368	2369	236A	236B	236C	236D	236E	236F	2370	2371
⍢	⍣	⍤	⍥	⍦	⍧	⍨	⍩	⍪	⍫	⍬	⍭	⍮	⍯	⍰	⍱
2372	2373	2374	2375	2376	2377	2378	2379	237A	237B	237C	237D	237E	237F	2380	2381
⍲	⍳	⍴	⍵	⍶	⍷	⍸	⍹	⍺	⍻	⍼	⍽	⍾	⍿	⎀	⎁
2382	2383	2384	2385	2386	2387	2388	2389	238A	238B	238C	238D	238E	238F	2390	2391
⎂	⎃	⎄	⎅	⎆	⎇	⎈	⎉	⎊	⎋	⎌	⎍	⎎	⎏	⎐	⎑
2392	2393	2394	2395	2396	2397	2398	2399	239A	239B	239C	239D	239E	239F	23A0	23A1
⎒	⎓	⎔	⎕	⎖	⎗	⎘	⎙	⎚	⎛	⎜	⎝	⎞	⎟	⎠	⎡
23A2	23A3	23A4	23A5	23A6	23A7	23A8	23A9	23AA	23AB	23AC	23AD	23B8	23B9	23BA	23BB
⎢	⎣	⎤	⎥	⎦	⎧	⎨	⎩	⎪	⎫	⎬	⎭	⎸	⎹	⎺	⎻
23BC	23BD	23BE	23BF	23C0	23C1	23C2	23C3	23C4	23C5	23C6	23C7	23C8	23C9	23CA	23CB
⎼	⎽	⎾	⎿	⏀	⏁	⏂	⏃	⏄	⏅	⏆	⏇	⏈	⏉	⏊	⏋
23CC	23CD	23CE	23CF	2423	2460	2461	2462	2463	2464	2465	2466	2467	2468	2469	246A
⏌	⏍	⏎	⏏	␣	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪

246B	246C	246D	246E	246F	2470	2471	2472	2473	24EA	24EB	24EC	24ED	24EE	24EF	24F0
⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳	⓪	⓫	⓬	⓭	⓮	⓯	⓰
24F1	24F2	24F3	24F4	24FF	2500	2502	250C	2510	2514	2518	251C	2524	252C	2534	2581
⓱	⓲	⓳	⓴	⓿	─	│	┌	┐	└	┘	├	┤	┬	┴	▁
2588	2592	25CB	25CC	2660	2661	2713	274D	2776	2777	2778	2779	277A	277B	277C	277D
■	▒	○	◌	♠	♡	✓	❍	❶	❷	❸	❹	❺	❻	❼	❽
277E	277F	27D0	27D1	27D2	27D3	27D4	27D5	27D6	27D7	27D8	27D9	27DA	27DB	27DC	27DD
❾	❿	⟐	⟑	⟒	⟓	⟔	⟕	⟖	⟗	⟘	⟙	⟚	⟛	⟜	⟝
27DE	27DF	27E0	27E1	27E2	27E3	27E4	27E5	27E6	27E7	27F0	27F1	27F2	27F3	27F4	27F5
⟞	⟟	⟠	⟡	⟢	⟣	⟤	⟥	⟦	⟧	⟰	⟱	⟲	⟳	⟴	⟵
27F6	27F7	27F8	27F9	27FA	27FB	27FC	27FD	27FE	27FF	2900	2901	2902	2903	2904	2905
⟶	⟷	⟸	⟹	⟺	⟻	⟼	⟽	⟾	⟿	⤀	⤁	⤂	⤃	⤄	⤅
2906	2907	2908	2909	290A	290B	290C	290D	290E	290F	2910	2911	2912	2913	2914	2915
⤆	⤇	⤈	⤉	⤊	⤋	⤌	⤍	⤎	⤏	⤐	⤑	⤒	⤓	⤔	⤕
2916	2917	2918	2919	291A	291B	291C	291D	291E	291F	2920	2921	2922	2923	2924	2925
⤖	⤗	⤘	⤙	⤚	⤛	⤜	⤝	⤞	⤟	⤠	⤡	⤢	⤣	⤤	⤥
2926	2927	2928	2929	292A	292B	292C	292D	292E	292F	2930	2931	2932	2933	2934	2935
⤦	⤧	⤨	⤩	⤪	⤫	⤬	⤭	⤮	⤯	⤰	⤱	⤲	⤳	⤴	⤵
2936	2937	2938	2939	293A	293B	293C	293D	293E	293F	2940	2941	2942	2943	2944	2945
⤶	⤷	⤸	⤹	⤺	⤻	⤼	⤽	⤾	⤿	⥀	⥁	⥂	⥃	⥄	⥅
2946	2947	2948	2949	294A	294B	294C	294D	294E	294F	2950	2951	2952	2953	2954	2955
⥆	⥇	⥈	⥉	⥊	⥋	⥌	⥍	⥎	⥏	⥐	⥑	⥒	⥓	⥔	⥕
2956	2957	2958	2959	295A	295B	295C	295D	295E	295F	2960	2961	2962	2963	2964	2965
⥖	⥗	⥘	⥙	⥚	⥛	⥜	⥝	⥞	⥟	⥠	⥡	⥢	⥣	⥤	⥥
2966	2967	2968	2969	296A	296B	296C	296D	296E	296F	2970	2971	2972	2973	2974	2975
⥦	⥧	⥨	⥩	⥪	⥫	⥬	⥭	⥮	⥯	⥰	⥱	⥲	⥳	⥴	⥵
2976	2977	2978	2979	297A	297B	297C	297D	297E	297F	2980	2981	2982	2983	2984	2985
⥶	⥷	⥸	⥹	⥺	⥻	⥼	⥽	⥾	⥿	⦀	⦁	⦂	⦃	⦄	⦅
2986	2987	2988	2989	298A	298B	298C	298D	298E	298F	2990	2991	2992	2993	2994	2995
⦆	⦇	⦈	⦉	⦊	⦋	⦌	⦍	⦎	⦏	⦐	⦑	⦒	⦓	⦔	⦕
2996	2997	2998	2999	299A	299B	299C	299D	299E	299F	29A0	29A1	29A2	29A3	29A4	29A5
⦖	⦗	⦘	⦙	⦚	⦛	⦜	⦝	⦞	⦟	⦠	⦡	⦢	⦣	⦤	⦥
29A6	29A7	29A8	29A9	29AA	29AB	29AC	29AD	29AE	29AF	29B0	29B1	29B2	29B3	29B4	29B5
⦦	⦧	⦨	⦩	⦪	⦫	⦬	⦭	⦮	⦯	⦰	⦱	⦲	⦳	⦴	⦵
29B6	29B7	29B8	29B9	29BA	29BB	29BC	29BD	29BE	29BF	29C0	29C1	29C2	29C3	29C4	29C5
⦶	⦷	⦸	⦹	⦺	⦻	⦼	⦽	⦾	⦿	⧀	⧁	⧂	⧃	⧄	⧅

29C6	29C7	29C8	29C9	29CA	29CB	29CC	29CD	29CE	29CF	29D0	29D1	29D2	29D3	29D4	29D5
⧆	⧇	⧈	⧉	⧊	⧋	⧌	⧍	⧎	⧏	⧐	⧑	⧒	⧓	⧔	⧕
29D6	29D7	29D8	29D9	29DA	29DB	29DC	29DD	29DE	29DF	29E0	29E1	29E2	29E3	29E4	29E5
⧖	⧗	⧘	⧙	⧚	⧛	⧜	⧝	⧞	⧟	⧠	⧡	⧢	⧣	⧤	⧥
29E6	29E7	29E8	29E9	29EA	29EB	29EC	29ED	29EE	29EF	29F0	29F1	29F2	29F3	29F4	29F5
⧦	⧧	⧨	⧩	⧪	⧫	⧬	⧭	⧮	⧯	⧰	⧱	⧲	⧳	⧴	⧵
29F6	29F7	29F8	29F9	29FA	29FB	29FE	29FF	2A00	2A01	2A02	2A03	2A04	2A05	2A06	2A07
⧶	⧷	⧸	⧹	⧺	⧻	⧾	⧿	⨀	⨁	⨂	⨃	⨄	⨅	⨆	⨇
2A08	2A09	2A0A	2A0B	2A0C	2A0D	2A0E	2A0F	2A10	2A11	2A12	2A13	2A14	2A15	2A16	2A17
⨈	⨉	⨊	⨋	⨌	⨍	⨎	⨏	⨐	⨑	⨒	⨓	⨔	⨕	⨖	⨗
2A18	2A19	2A1A	2A1B	2A1C	2A1D	2A1E	2A1F	2A20	2A21	2A22	2A23	2A24	2A25	2A26	2A27
⨘	⨙	⨚	⨛	⨜	⨝	⨞	⨟	⨠	⨡	⨢	⨣	⨤	⨥	⨦	⨧
2A28	2A29	2A2A	2A2B	2A2C	2A2D	2A2E	2A2F	2A30	2A31	2A32	2A33	2A34	2A35	2A36	2A37
⨨	⨩	⨪	⨫	⨬	⨭	⨮	⨯	⨰	⨱	⨲	⨳	⨴	⨵	⨶	⨷
2A38	2A39	2A3A	2A3B	2A3C	2A3D	2A3E	2A3F	2A40	2A41	2A42	2A43	2A44	2A45	2A46	2A47
⨸	⨹	⨺	⨻	⨼	⨽	⨾	⨿	⩀	⩁	⩂	⩃	⩄	⩅	⩆	⩇
2A48	2A49	2A4A	2A4B	2A4C	2A4D	2A4E	2A4F	2A50	2A51	2A52	2A53	2A54	2A55	2A56	2A57
⩈	⩉	⩊	⩋	⩌	⩍	⩎	⩏	⩐	⩑	⩒	⩓	⩔	⩕	⩖	⩗
2A58	2A59	2A5A	2A5B	2A5C	2A5D	2A5E	2A5F	2A60	2A61	2A62	2A63	2A64	2A65	2A66	2A67
⩘	⩙	⩚	⩛	⩜	⩝	⩞	⩟	⩠	⩡	⩢	⩣	⩤	⩥	⩦	⩧
2A68	2A69	2A6A	2A6B	2A6C	2A6D	2A6E	2A6F	2A70	2A71	2A72	2A73	2A74	2A75	2A76	2A77
⩨	⩩	⩪	⩫	⩬	⩭	⩮	⩯	⩰	⩱	⩲	⩳	⩴	⩵	⩶	⩷
2A78	2A79	2A7A	2A7B	2A7C	2A7D	2A7E	2A7F	2A80	2A81	2A82	2A83	2A84	2A85	2A86	2A87
⩸	⩹	⩺	⩻	⩼	⩽	⩾	⩿	⪀	⪁	⪂	⪃	⪄	⪅	⪆	⪇
2A88	2A89	2A8A	2A8B	2A8C	2A8D	2A8E	2A8F	2A90	2A91	2A92	2A93	2A94	2A95	2A96	2A97
⪈	⪉	⪊	⪋	⪌	⪍	⪎	⪏	⪐	⪑	⪒	⪓	⪔	⪕	⪖	⪗
2A98	2A99	2A9A	2A9B	2A9C	2A9D	2A9E	2A9F	2AA0	2AA1	2AA2	2AA3	2AA4	2AA5	2AA6	2AA7
⪘	⪙	⪚	⪛	⪜	⪝	⪞	⪟	⪠	⪡	⪢	⪣	⪤	⪥	⪦	⪧
2AA8	2AA9	2AAA	2AAB	2AAC	2AAD	2AAE	2AAF	2AB0	2AB1	2AB2	2AB3	2AB4	2AB5	2AB6	2AB7
⪨	⪩	⪪	⪫	⪬	⪭	⪮	⪯	⪰	⪱	⪲	⪳	⪴	⪵	⪶	⪷
2AB8	2AB9	2ABA	2ABB	2ABC	2ABD	2ABE	2ABF	2AC0	2AC1	2AC2	2AC3	2AC4	2AC5	2AC6	2AC7
⪸	⪹	⪺	⪻	⪼	⪽	⪾	⪿	⫀	⫁	⫂	⫃	⫄	⫅	⫆	⫇
2AC8	2AC9	2ACA	2ACB	2ACC	2ACD	2ACE	2ACF	2AD0	2AD1	2AD2	2AD3	2AD4	2AD5	2AD6	2AD7
⫈	⫉	⫊	⫋	⫌	⫍	⫎	⫏	⫐	⫑	⫒	⫓	⫔	⫕	⫖	⫗
2AD8	2AD9	2ADA	2ADB	2ADC	2ADD	2ADE	2ADF	2AE0	2AE1	2AE2	2AE3	2AE4	2AE5	2AE6	2AE7
⫘	⫙	⫚	⫛	⫝̸	⫝	⫞	⫟	⫠	⫡	⫢	⫣	⫤	⫥	⫦	⫧

2AE8	2AE9	2AEA	2AEB	2AEC	2AED	2AEE	2AEF	2AF0	2AF1	2AF2	2AF3	2AF4	2AF5	2AF6	2AF7	F21D	F21E	F21F	F220	F221	F222	F223	F224	F225	F226	F227	F228	F229	F22A	F22B	F22C
⫠	⫩	⫪	⫫	⫬	⫭	⫮	⫯	⫰	⫱	⫲	⫳	⫴	⫵	⫶	⫷	ˈ	ʔ	Ł	ɐ	Ɐ	ɦ	Ɦ	ɧ	ç	ɖ	ʃ	ɡ	ƙ	ɻ	ɱ	ɳ
2AF8	2AF9	2AFA	2AFB	2AFC	2AFD	2AFE	2AFF	2B04	2B06	2B07	2B0C	2B0D	2C60	2C60	2C61	F22D	F22E	F22F	F230	F231	F232	F233	F234	F235	F236	F237	F238	F239	F23A	F23B	F23C
⫸	⫷	⫸	⫽	⫼	⫽	▯	▯	⇔	⬆	⬇	↔	↕	Ⱡ	Ⱡ	ⱡ	ɒ	ɾ	ʂ	ʃ	ʏ	ɣ	ʒ	Ʒ	ʒ	ą	ɑ	ɐ	ɛ	ɜ	ə	į
2C61	2C62	2C62	2C63	2C63	2C64	2C64	2C65	2C66	2C67	2C67	2C68	2C68	2C69	2C69	2C6A	F23D	F23E	F23F	F240	F241	F242	F243	F244	F245	F246	F247	F248	F249	F24A	F24B	F24C
ⱡ	Ł	Ł	Ᵽ	Ᵽ	Ɽ	Ɽ	ⱥ	ȿ	Ⱨ	Ⱨ	ⱨ	ⱨ	Ⱪ	Ⱪ	ⱪ	ɕ	ɥ	ʉ	ɗ	ɸ	Ł	ɣ	Ɣ	ɛ	ɡ	@	Ⓐ	ƀ	đ	ɬ	ɯ
2C6A	2C6B	2C6B	2C6C	2C6C	2C6D	2C6E	2C6F	2C70	2C71	2C72	2C73	2C74	2C74	2C75	2C76	F24D	F24E	F24F	F250	F251	F252	F253	F254	F255	F256	F257	F258	F259	F25A	F25B	F25C
ⱪ	Ⱬ	Ⱬ	ⱬ	ⱬ	Ɑ	Ɱ	Ⱶ	Ɒ	ⱱ	Ⱳ	ⱳ	ⱴ	ⱴ	Ⱶ	ⱶ	ɲ	ɸ	ɾ	ɾ	ʂ	ŧ	ƶ	ɨ	ʉ	ʂ	ɀ	ɘ	ɵ	ɧ	Ɱ	Ᵽ
2C77	2C78	2C79	2C7A	2C7B	2C7C	2C7D	2E17	3014	3015	3016	3017	A64C	A64D	A700	A701	F25D	F25E	F25F	F260	F261	F262	F263	F264	F265	F266	F267	F268	F269	F26A	F26B	F320
ⱷ	ɐ	ɭ	ɷ	Ɜ	ʲ	ⱽ	⸗	〔	〕	⟦	⟧	Ꝍ	ꝍ	ˠ	ˤ	ŧ	ʋ	ⱱ	ʲ	ɟ	ƙ	Ƙ	ʐ	Ʒ	ɬ	ɤ	ɬ	ɮ	ˈ	Ч	Ɣ
A702	A703	A704	A705	A706	A707	A708	A709	A70A	A70B	A70C	A70D	A70E	A70F	A710	A711	F321	F322	F323	F324	F325	F326	F327	F328	F329	F32A	F32B	F32C	F32D	FB00	FB03	FB04
˂	˃	˄	˅	˄	˅	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	ɻ	ɺ	ɺ	Ɣ	ɣ	Ƒ	ƒ	Ɤ	ɤ	ɛ	ɛ	ɧ	ɧ	ff	ffi	ffl
A712	A713	A714	A715	A716	A717	A717	A718	A718	A719	A719	A71A	A71A	A71B	A71C	A71D	FE00	FE01	FE02	FE03	FE04	FE05	FE06	FE07	FE08	FE09	FE0A	FE0B	FE0C	FE0D	FE0E	FE0F
˥	˦	˦	˦	˩	˖	˖	˗	˗	˭	˭	˩	˩	↑	↓	ꜜ																
A71E	A71F	A720	A721	A722	A723	A724	A725	A726	A727	A728	A729	A72A	A72B	A72C	A72D	FE20	FE21	FE22	FE23	FEFF	FFF9	FFFA	FFFB	FFFC	FFFD	D510	D513	0000	***	***	***
ˡ	ˡ	˧	˧	˱	˱	˂	˂	Ⱨ	ⱨ	Ⱬ	ⱬ	ɛ	ɛ	ɣ	ɣ	⁀	⁀	⁀	⁀						□						
A72E	A72F	A730	A731	A732	A733	A734	A735	A740	A741	A74E	A74F	A788	A789	A78A	A78B	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ɣ	ɣ	ꜰ	ꜱ	Ꜳ	ꜳ	Ꜵ	ꜵ	Ꝁ	ꝁ	Ꝏ	ꝏ	ˆ	ː	˭	ˈ																
A78C	A78D	A78E	A7FB	A7FC	A7FD	A7FE	A7FF	F130	F131	F132	F133	F134	F135	F170	F171	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ˈ	Ч	ɬ	Ꜧ	ɋ	Ɯ	Ɪ	Ʌ	ˌ	ˌ	ˍ	ˍ	˩	˩	˷	ˉ																
F172	F173	F174	F175	F176	F178	F179	F17A	F17B	F180	F181	F182	F18B	F195	F196	F197	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
˷	˷	˷	˷	‿	ʳ	~	¨	≈	m	ɲ	ŋ	ɜ	˵	ˈ	ˊ																
F198	F199	F19A	F19B	F19C	F19D	F19E	F19F	F1A0	F1A1	F1A2	F1A3	F1A4	F1A5	F1A6	F1A7	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
˩	˖	˗	˭	↑	↓	!	¡	ɒ	æ	ɕ	ɘ	ɞ	f	g	ɨ																
F1A8	F1A9	F1AA	F1AB	F1AC	F1AD	F1AE	F1AF	F1B0	F1B1	F1B2	F1B3	F1B4	F1B5	F1B6	F1B7	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ɪ	ɟ	ɭ	ø	ɵ	œ	œ	ʃ	ɥ	ʊ	ɯ	ʌ	ʏ	ɤ	z	ʐ																
F1B8	F1B9	F1BA	F1BB	F1BC	F1BD	F1BE	F1BF	F1C0	F1C1	F1C2	F1C3	F1C4	F1C5	F1C6	F1C7	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ʒ	c	ð	ɥ	ɦ	ɟ	ɻ	ʟ	ɱ	ɳ	ɴ	ɸ	ʂ	ʈ	ʊ	ʐ																
F1C8	F1C9	F1CA	F1CB	F1CC	F1CD	F1CE	F1D0	F1D1	F1D2	F1D3	F1D4	F1D5	F1D6	F1D7	F1D8	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ʔ	θ	ɩ	ɨ	ᴜ	ʉ	ʏ	⎡	⎢	⎢	⎢	⎣	˙ǀ	˙ǀ	˙ǀ	˙ǀ																
F1D9	F1DA	F1DB	F1DC	F1DD	F1DE	F1DF	F1E0	F1E1	F1E2	F1E3	F1E4	F1E5	F1E6	F1E7	F1E8	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˙ǀ	˂	˂	˃	˃	˂	˂	˃	˃	ˆ	ˇ																
F1E9	F1EA	F1F1	F1F2	F1F3	F1F4	F1F5	F1F6	F1F7	F1F8	F1F9	F208	F209	F20A	F20B	F20C	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
ː	˭	_	_	_	_	_	_	_	_	_	ɑ	ɒ	ʙ	ɕ	ɖ																
F20D	F20E	F20F	F210	F211	F212	F213	F214	F215	F216	F217	F218	F219	F21A	F21B	F21C	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Ð	ɬ	Ł	ᵽ	ɋ	ɋ	ɾ	ʀ	ʀ	ʃ	ʒ	ʉ	ʌ	ʍ	ʍ	ʒ																

JOINT Hamburgefonts Charis SIL

JOINT Hamburgefonts

JOINT Hamburgefonts Cambria

In traditional typography, text is composed to create a readable, coherent, and visually satisfying whole that works invisibly, without the awareness of the reader. Even distribution of typeset material, with a minimum of distractions and anomalies, is aimed at producing clarity and transparency. Choice of typeface(s) is the primary aspect of text typography—prose fiction, non-fiction, editorial, educational, religious, scientific, spiritual and commercial writing all have differing characteristics and requirements of appropriate typefaces and fonts. For historic material established text typefaces are frequently chosen according to a scheme of historical genre acquired by a long process of accretion, with considerable overlap between

In traditional typography, text is composed to create a readable, coherent, and visually satisfying whole that works invisibly, without the awareness of the reader. Even distribution of typeset material, with a minimum of distractions and anomalies, is aimed at producing clarity and transparency. Choice of typeface(s) is the primary aspect of text typography—prose fiction, non-fiction, editorial, educational, religious, scientific, spiritual and commercial writing all have differing characteristics and requirements of appropriate typefaces and fonts. For historic material established text typefaces are frequently chosen according to a scheme of historical genre acquired by a long process of accretion, with considerable overlap between

Conclusion

adam@twardoch.com
wolinski@ipipan.waw.pl