





# Journal of Language Modelling

VOLUME 12 ISSUE 1  
JUNE 2024

## Articles

The expected sum of edge lengths in planar linearizations of trees 1  
*Lluís Alemany-Puig, Ramon Ferrer-i-Cancho*

Detecting inflectional patterns for Croatian verb stems  
using class activation mapping 43  
*Domagoj Ševerdija, Rebeka Čorić, Marko Orešković, Lucian Šošić*

Control, inner topicalisation, and focus fronting in Mandarin Chinese:  
modelling in parallel constraint-based grammatical architecture 69  
*Chit-Fung Lam*

On German verb sense disambiguation:  
A three-part approach based on linking a sense inventory  
(GermaNet) to a corpus through annotation (TGVCorp)  
and using the corpus to train a VSD classifier (TTvSense) 155  
*Dominik Mattern, Wahed Hemati, Andy Lücking, Alexander Mehler*

## Tools and resources

QRGS – Question Responses Generation via crowdsourcing 213  
*Paweł Łupkowski, Jonathan Ginzburg, Ewelina Chmurska,  
Adrianna Płatosz, Aleksandra Kwiecień, Barbara Adamska,  
Magdalena Szkalej*



JOURNAL OF  
LANGUAGE MODELLING

ISSN 2299-8470 (electronic version)

ISSN 2299-856X (printed version)

<http://jlm.ipipan.waw.pl/>

MANAGING EDITOR

*Adam Przepiórkowski* IPI PAN

SECTION EDITORS

*Elżbieta Hajnicz* IPI PAN

*Małgorzata Marciniak* IPI PAN

*Agnieszka Mykowiecka* IPI PAN

*Marcin Woliński* IPI PAN

STATISTICS EDITOR

*Łukasz Dębowski* IPI PAN



Published by IPI PAN


Institute of Computer Science, Polish Academy of Sciences  
ul. Jana Kazimierza 5, 01-248 Warszawa, Poland

Circulation: 50 + print on demand

Layout designed by Adam Twardoch.

Typeset in X<sub>Y</sub>L<sup>A</sup>T<sub>E</sub>X using the typefaces: *Playfair*  
by Claus Eggers Sørensen, *Charis SIL* by SIL International,  
*JLM monogram* by Łukasz Dziedzic.

*All content is licensed under  
the Creative Commons Attribution 4.0 International License.*

 <http://creativecommons.org/licenses/by/4.0/>

## EDITORIAL BOARD

*Steven Abney* University of Michigan, USA

*Ash Asudeh* University of Rochester, USA

*Igor Boguslavsky* Technical University of Madrid, SPAIN

*Paul Boersma* University of Amsterdam, THE NETHERLANDS

*Olivier Bonami* Université Paris Cité,  
Laboratoire de linguistique formelle, CNRS, FRANCE

*Robert D. Borsley* Professor Emeritus, University of Essex;  
Honorary Professor, Bangor University, UNITED KINGDOM

*António Branco* University of Lisbon, PORTUGAL

*David Chiang* University of Notre Dame, USA

*Dan Cristea* University of Iași, ROMANIA

*Berthold Crysmann* Université Paris Cité,  
Laboratoire de linguistique formelle, CNRS, FRANCE

*Jan Daciuk* Gdańsk University of Technology, POLAND

*Lukasz Dębowski* Institute of Computer Science,  
Polish Academy of Sciences, POLAND

*Mary Dalrymple* Professor Emerita, University of Oxford, UNITED KINGDOM

*Anette Frank* Universität Heidelberg, GERMANY

*Claire Gardent* LORIA, CNRS and Université de Lorraine, FRANCE

*Jonathan Ginzburg* Université Paris Cité, Laboratoire de linguistique  
formelle, CNRS; Laboratoire d'Excellence LabEx-EFLt, FRANCE

*Thomas Graf* Stony Brook University, UNITED STATES

*Stefan Th. Gries* University of California, Santa Barbara, USA;  
Justus Liebig University Giessen, GERMANY

*Adam Jardine* Rutgers Department of Linguistics, UNITED STATES

*Heiki-Jaan Kaalep* University of Tartu, ESTONIA

*Laura Kallmeyer* Heinrich-Heine-Universität Düsseldorf, GERMANY

*Jong-Bok Kim* Kyung Hee University, Seoul, KOREA

*Kimmo Koskenniemi* Professor Emeritus, University of Helsinki, FINLAND

*Jonas Kuhn* Universität Stuttgart, GERMANY

*Alessandro Lenci* University of Pisa, ITALY

*John J. Lowe* University of Oxford, UNITED KINGDOM

*Ján Mačutek* Comenius University, Bratislava, SLOVAKIA

*Igor Mel'čuk* Professor Emeritus, University of Montreal, CANADA

*Richard Moot* CNRS, LIRMM, University of Montpellier, FRANCE

*Glyn Morrill* Technical University of Catalonia, Barcelona, SPAIN

*Stefan Müller* Humboldt Universität zu Berlin, GERMANY

*Mark-Jan Nederhof* University of St Andrews, UNITED KINGDOM

*Petya Osenova* Sofia University, BULGARIA

*David Pesetsky* Massachusetts Institute of Technology, USA

*Maciej Piasecki* Wrocław University of Science and Technology, POLAND

*Christopher Potts* Stanford University, USA

*Agata Savary* University of Paris-Saclay, FRANCE

*Sabine Schulte im Walde* Universität Stuttgart, GERMANY

*Stuart M. Shieber* Harvard University, USA

*Mark Steedman* University of Edinburgh, UNITED KINGDOM

*Stan Szpakowicz* Professor Emeritus, University of Ottawa, CANADA

*Shravan Vasishth* Universität Potsdam, GERMANY

*Aline Villavicencio* Institute for Data Science and Artificial Intelligence  
University of Exeter; University of Sheffield, UNITED KINGDOM

*Veronika Vincze* University of Szeged, HUNGARY

*Shuly Wintner* University of Haifa, ISRAEL

*Zdeněk Žabokrtský* Charles University in Prague, CZECH REPUBLIC

# The expected sum of edge lengths in planar linearizations of trees

*Lluís Alemany-Puig and Ramon Ferrer-i-Cancho*  
Universitat Politècnica de Catalunya (UPC)

## ABSTRACT

Dependency trees have proven to be a very successful model to represent the syntactic structure of sentences of human languages. In these structures, vertices are words and edges connect syntactically-dependent words. The tendency of these dependencies to be short has been demonstrated using random baselines for the sum of the lengths of the edges or their variants. A ubiquitous baseline is the expected sum in projective orderings (wherein edges do not cross and the root word of the sentence is not covered by any edge), that can be computed in time  $O(n)$ . Here we focus on a weaker formal constraint, namely planarity. In the theoretical domain, we present a characterization of planarity that, given a sentence, yields either the number of planar permutations or an efficient algorithm to generate uniformly random planar permutations of the words. We also show the relationship between the expected sum in planar arrangements and the expected sum in projective arrangements. In the domain of applications, we derive a  $O(n)$ -time algorithm to calculate the expected value of the sum of edge lengths. We also apply this research to a parallel corpus and find that the gap between actual dependency distance and the random baseline reduces as the strength of the formal constraint on dependency structures increases, suggesting that formal constraints absorb part of the dependency distance minimization effect. Our research paves the way for replicating past research on dependency distance minimization using random planar linearizations as random baseline.

*Keywords:*  
*dependency*  
*grammar,*  
*projectivity,*  
*planarity,*  
*syntactic*  
*dependency*  
*distance*  
*minimization*

The structure of a natural language sentence can be represented as a (labelled) graph indicating the syntactic relationships between words together with the encoding of the words' order. In such a graph, the edge labels indicate the type of syntactic relationship between the words. Such a combination of graph and linear ordering, as in Figure 1, is known as syntactic dependency structure (Nivre 2006). When the graph is (1) *well-formed*, namely, the graph is weakly connected, (2) is *acyclic*, that is, there are no cycles in the graph, (3) is *single-headed*, that is, every node has a single head (except for the root node), and (4) there is only one root node (one node with no head) in the graph, then it is called a syntactic dependency tree (Nivre 2006). There exist *formal constraints* that are often imposed on dependency structures. One such constraint is projectivity: a dependency structure is projective if, for every vertex  $v$ , all vertices reachable from  $v$  in the underlying graph form a continuous substring within the sentence (Kuhlmann and Nivre 2006). Projectivity implies that (1) the root word of the sentence (the root of the underlying syntactic dependency structure) is never covered (as in Figure 1(a)) and (2) planarity, namely absence of edge crossings (Figure 1 (a) and (b)). Indeed planarity is another constraint that generalizes projectivity by allowing the root to be covered by one or more edges (as in Figure 1(b)). Figure 1(c) shows a sentence that is neither projective nor planar.

In this article, we study statistical properties of syntactic dependency structures under the planarity constraint. Such structures are represented in this article as a pair consisting of a (free or rooted) tree and a linear arrangement of its vertices. Free trees are denoted as  $T = (V, E)$ , and rooted trees as  $T^r = (V, E; r)$ , where  $V$  is the set of vertices,  $E$  the set of edges, and  $r \in V$  denotes the root vertex. Unless stated otherwise  $n = |V|$ , that is,  $n$  denotes the number of vertices which is equal to the number of words in the sentence. A linear arrangement  $\pi$  (also called *embedding*) of a tree is a (bijective) function ( $\pi : V \rightarrow \{1, \dots, n\}$ ) that maps every vertex  $u$  of a tree to a unique position in  $\{1, \dots, n\}$ , which is denoted by  $\pi(u)$ .

Projectivity, as well as planarity, can be alternatively defined on linear arrangements using the concept of edge crossing. We say that



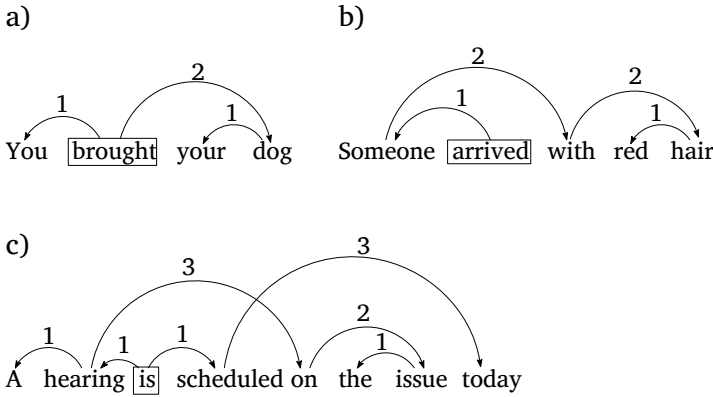


Figure 1: Examples of sentences with their syntactic dependency structures; arc labels indicate dependency distance (in words) between linked words. The rectangles denote the root word in each sentence. a) A projective dependency tree (adapted from Groß and Osborne 2009). b) Planar (but not projective) syntactic dependency structure (adapted from Groß and Osborne 2009). c) Non-projective and non-planar syntactic dependency structure (adapted from Nivre 2009)

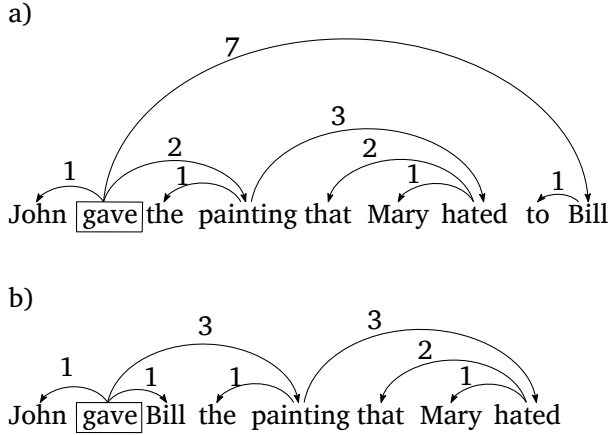
any two (undirected) edges  $\{s, t\}$ ,  $\{u, v\}$  cross if the positions of their vertices interleave. More formally, assume, without loss of generality, that  $\pi(s) < \pi(t)$ ,  $\pi(u) < \pi(v)$  and  $\pi(s) < \pi(u)$ . Then, edges  $\{s, t\}$ ,  $\{u, v\}$  cross in the linear ordering defined by  $\pi$  if  $\pi(s) < \pi(u) < \pi(t) < \pi(v)$ .<sup>1</sup> We denote the total number of edge crossings in an arrangement  $\pi$  as  $C_\pi(T)$ . Then, an arrangement  $\pi$  of a rooted tree  $T^r$  is *planar* if  $C_\pi(T^r) = 0$  and is *projective* if (a) it is planar and (b) the root of the tree is not covered, that is, there is no edge  $\{s, t\}$  such that  $\pi(s) < \pi(r) < \pi(t)$  or  $\pi(t) < \pi(r) < \pi(s)$ . Planarity is a relaxation of projectivity where the root can be covered (Sleator and Temperley 1993; Kuhlmann and Nivre 2006). Planar arrangements are also known in the literature as *one-page book embeddings* (Bernhart and Kainen 1979).

In this article, the main object of study is the expectation of the sum of edge lengths (or syntactic dependency distances) in planar arrangements of free trees. The length of an edge connecting two syntactically-related words, also known as dependency distance, is usually<sup>2</sup> defined as the number of intervening words between  $u$  and  $v$

<sup>1</sup>Notice that this notion of crossing does not depend on edge orientation.

<sup>2</sup>Another popular definition is  $\delta_{uv}(\pi) = |\pi(u) - \pi(v)| - 1$  (Liu *et al.* 2017).

Figure 2:  
 Examples of sentences with  
 their syntactic dependency  
 structures; arc labels  
 indicate dependency  
 distance. The rectangles  
 denote the root word in  
 each sentence. Examples  
 adapted from Morrill 2000.  
 The sum of edge lengths  
 are  $D = 18$  for (a) and  
 $D = 12$  for (b)



in the sentence plus 1 (Figure 1). It is defined mathematically as

$$\delta_{uv}(\pi) = |\pi(u) - \pi(v)|.$$

We define the total sum of edge lengths in  $\pi$  as

$$(1) \quad D_{\pi}(T) = \sum_{uv \in E} \delta_{uv}(\pi).$$

Close attention has been paid to this metric in modern linguistic research since its causal relationship with cognitive cost was first put forward, to the best of our knowledge, by Hudson 1995. The main causal argument is that the longer the dependency, the greater the memory burden arising from decay of activation and interference (Hudson 1995; Liu *et al.* 2017). A number of studies have exposed the general tendency in languages to reduce  $D$ , the total sum of edge lengths, a reflection of a potentially universal cognitive force known as the Dependency Distance Minimization principle (DDm) (Ferrer-i-Cancho 2004; Liu 2008; Futrell *et al.* 2015; Liu *et al.* 2017; Ferrer-i-Cancho *et al.* 2022). As an example of such cognitive cost, consider the sentences in Figures 2(a) and 2(b): it is not surprising that the latter is preferred over the former due to smaller total sum of edge lengths (Morrill 2000), the former's being  $D = 18$  and the latter's being  $D = 12$ .

Statistical evidence of the DDm principle has been provided showing that dependency distances are smaller than expected by chance in syntactic dependency treebanks (Ferrer-i-Cancho 2004; Liu 2008; Park and Levy 2009; Gildea and Temperley 2010; Futrell *et al.* 2015; Liu

*et al.* 2017; Ferrer-i-Cancho *et al.* 2022; Kramer 2021). Typically, the random baseline is defined as a random shuffling of the words of a sentence. To the best of our knowledge, the first known instance of such an approach was done by Ferrer-i-Cancho 2004, who established the DDm principle by comparing the average real  $D(T)$  of sentences against the corresponding expected value in a uniformly random permutation of sentences' words. More formally, Ferrer-i-Cancho 2004 calculated the expected value of  $D(T)$  when the words of the sentence are shuffled uniformly at random (u.a.r.), that is, when all  $n!$  permutations are equally likely. This value is denoted here as  $\mathbb{E}[D(T)]$ . Ferrer-i-Cancho 2004 found that

$$(2) \quad \mathbb{E}[D(T)] = \frac{n^2 - 1}{3}.$$

In spite of the simplicity of Equation 2, the majority of researchers have used as random baseline the expected sum of edge lengths conditioned to projective arrangements (Temperley 2008; Park and Levy 2009; Gildea and Temperley 2010; Futrell *et al.* 2015; Kramer 2021) which we denote here as  $\mathbb{E}_{\text{pr}}[D(T^r)]$ . However, this baseline has been computed approximately via random sampling of projective arrangements. For these reasons, a formula to calculate the exact value of  $\mathbb{E}_{\text{pr}}[D(T^r)]$  in linear time was derived by Alemany-Puig and Ferrer-i-Cancho 2022

$$(3) \quad \mathbb{E}_{\text{pr}}[D(T^r)] = \frac{1}{6} \sum_{u \in V} s_r(u)(2d_r(u) + 1) - \frac{1}{6},$$

where  $s_r(u)$  denotes the size (in vertices) of the subtree of  $T^r$  rooted at  $u$ , and  $d_r(u)$  is the out-degree of  $u$  in  $T^r$ . In spite of its extensive use, the projective random baseline has some limitations. First, the percentage of non-projective sentences in languages ranges between 18.2 and 26.4 (Gómez-Rodríguez 2016) or between 6.8 and 36.4 (Gómez-Rodríguez and g 2010) (see also Havelka 2007). The limited coverage of projectivity raises the question if the projective baseline should be used for sentences that are not projective as it is customary in research on dependency distance minimization. In addition, projectivity *per se* implies a reduction in dependency distances, which raises the question if that rather strong constraint may mask the effect of the dependency distance minimization principle under investigation (Gómez-

Rodríguez *et al.* 2022). Here we aim to make a step forward by considering planarity, a generalization of projectivity, so as to increase the coverage of real sentences and reduce the bias towards dependency minimization in the random baseline. The percentage of non-planar sentences in languages ranges between 14.3 and 20.0 (Ferrer-i-Cancho *et al.* 2018) or between 5.3 and 31 (Gómez-Rodríguez and g 2010). The latter range is consistent with earlier estimates (Havelka 2007).

This article is part of a research program on the statistical properties of  $D(T)$  under constraints on the possible linear arrangements (Ferrer-i-Cancho 2019; Alemany-Puig *et al.* 2022; Alemany-Puig and Ferrer-i-Cancho 2022). The remainder of the article is divided into two main parts: theory (Section 2) and applications (Section 3).

The theory part (Section 2) is structured as follows. In Section 2.1, we introduce notation used throughout that part. In Section 2.2, we first present a characterization of planar arrangements so as to identify their underlying structure, which we apply to count their number for a given free tree, and later on in Section 2.3, to generate them u.a.r. by means of a novel  $O(n)$ -time algorithm. In Section 2.4, we use said characterization to prove the main result of the article, namely that expectation of  $D(T)$  in planar arrangements can be calculated from the expectation of projective arrangements, as the following theorem indicates.

**THEOREM 1**      *Given a free tree  $T = (V, E)$ ,*

$$(4) \quad \mathbb{E}_{\text{pl}}[D(T)] = \frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}}^{\circ}[D(T^u)]$$

$$(5) \quad = \frac{(n-1)(n-2)}{6n} + \frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}}[D(T^u)],$$

where  $\mathbb{E}_{\text{pr}}^{\circ}[D(T^u)]$  is the expected value of  $D(T^u)$  in uniformly random projective arrangements  $\pi$  of  $T^u$  such that  $\pi(u) = 1$  and  $\mathbb{E}_{\text{pr}}[D(T^u)]$  (Equation 3) is the expected value of  $D(T^u)$  in uniformly random projective arrangements of  $T^u$ , the free tree  $T$  rooted at  $u$ .

Table 1 summarizes the theoretical results obtained in previous articles and those presented in this article.

The applications part (Section 3) is structured as follows. In Section 3.1, we apply Theorem 1 to derive a  $O(n)$ -time algorithm to calculate  $\mathbb{E}_{\text{pl}}[D(T)]$ . Since Alemany-Puig and Ferrer-i-Cancho 2022

Table 1: Summary of the main mathematical results for increasing constraints on linear orders. Results for the unconstrained and projective cases are borrowed from previous research (Ferrer-i-Cancho 2004 and Alemany-Puig and Ferrer-i-Cancho 2022, respectively). Results for the planar case are a contribution of this article.  $\mathbf{N}_{\text{pr}}(T^r)$ ,  $\mathbf{N}_{\text{pl}}(T)$  and  $\mathbf{N}(T)$  denote the number of distinct projective, planar and unconstrained linear arrangements, respectively, of a rooted tree  $T^r$  or of a free tree  $T$ .  $\mathbb{E}_{\text{pr}}[\delta_{uv}]$ ,  $\mathbb{E}_{\text{pl}}[\delta_{uv}]$  and  $\mathbb{E}[\delta_{uv}]$  denote the expected length of an edge in random linear arrangement for the projective, planar and unconstrained cases, respectively.  $\mathbb{E}_{\text{pr}}[\delta_{uv} | s]$  is the expected value of  $\delta_{uv}$  conditioned to having vertex  $s$  as root of the tree. In  $\mathbb{E}_{\text{pr}}[\delta_{uv}]$  the root is vertex  $r$

Unconstrained ( $T$ )	$\mathbf{N}(T)$	$n!$
	$\mathbb{E}[\delta_{uv}]$	$\frac{n+1}{3}$
	$\mathbb{E}[D(T)]$	$\frac{n^2-1}{3}$
Planar ( $T$ )	$\mathbf{N}_{\text{pl}}(T)$	$n \prod_{u \in V} d(u)!$
	$\mathbb{E}_{\text{pl}}[\delta_{uv}]$	$1 + \frac{1}{n} \sum_{s \in V \setminus \{u,v\}} \mathbb{E}_{\text{pr}}[\delta_{uv}   s]$
	$\mathbb{E}_{\text{pl}}[D(T)]$	$\frac{(n-1)(n-2)}{6n} + \frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}}[D(T^u)]$
Projective ( $T^r$ )	$\mathbf{N}_{\text{pr}}(T^r)$	$\prod_{u \in V} (d_r(u) + 1)!$
	$\mathbb{E}_{\text{pr}}[\delta_{uv}]$	$\frac{1}{6} (2s_r(u) + s_r(v) + 1)$
	$\mathbb{E}_{\text{pr}}[D(T^r)]$	$\frac{1}{6} \left( -1 + \sum_{v \in V} s_r(v) (2d_r(v) + 1) \right)$

showed that  $\mathbb{E}_{\text{pr}}[D(T^r)]$  can be evaluated in time  $O(n)$ , Equation 5 naturally leads to a  $O(n^2)$ -time algorithm if it is evaluated ‘as is’. However, we devise a  $O(n)$ -time algorithm to calculate  $\mathbb{E}_{\text{pl}}[D(T)]$ . In Section 3.2, we apply this and previous research on the projective case (Alemany-Puig and Ferrer-i-Cancho 2022) to a parallel syntactic dependency treebank. We find that the gap between the actual dependency distance and that of the random baseline reduces as the strength of the formal constraint on dependency structures chosen for the ran-

dom baseline increases, suggesting that formal constraints absorb part of the dependency distance minimization effect.

Finally, in Section 4, we review all the findings and make suggestions for future research.

From this point onwards, the article is organized to ease reading by readers of distinct profiles. Readers interested in the analysis of syntactic dependency treebanks can jump directly to Section 3.2. Readers interested in the algorithm for computing  $\mathbb{E}_{\text{pl}}[D(T)]$  can jump directly to Section 3.1, after reading Section 2.1. Readers whose primary interest is applying the algorithms have ready-to-use code: both methods to generate planar arrangements (Section 2.3) and the  $O(n)$ -time calculation of  $\mathbb{E}_{\text{pl}}[D(T)]$  (Section 3.1) are freely available in the Linear Arrangement Library<sup>3</sup> (Alemany-Puig *et al.* 2021).

## 2

## THEORY

### 2.1

#### *Definitions and notation*

We use  $u, v, w, z$  to denote vertices,  $r$  to always denote a root vertex, and  $i, j, k, p, q$  to denote integers. The edges of a free tree are undirected, and denoted as  $\{u, v\} = uv$ ; those of rooted trees are directed, denoted as  $(u, v)$ , and oriented away from  $r$  towards the leaves.

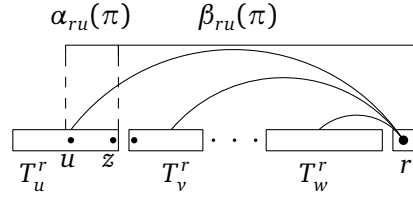
Let  $\Gamma(u)$  denote the set of neighbors of  $u \in V$  in the free tree  $T$ , and let  $\Gamma_r(u)$  denote the out neighbors (also, children) of  $u \in V$  in  $T^r$ . Notice that,  $\Gamma_r(u) \subseteq \Gamma(u)$  with equality if, and only if  $u = r$ . Let  $d_r(u) = |\Gamma_r(u)|$  denote the out-degree of vertex  $u$  of a rooted tree  $T^r$ , and let  $d(u) = |\Gamma(u)|$  denote the degree of  $u$  in a free tree  $T$ . Notice that  $d_r(u) = d(u) - 1$  when  $u \neq r$  and  $d_r(r) = d(r)$ . Furthermore, we denote the subtree rooted at  $v$  with respect to root  $u$  as  $T_v^u$  (obviously  $T_r^r = T^r$ ), and its size as  $s_u(v) = |V(T_v^u)|$  (Figure 3). We call this *directional size* (Hochberg and Stallmann 2003; Alemany-Puig *et al.* 2022). Note that  $s_v(u) + s_u(v) = n$  for any  $uv \in E$ .

---

<sup>3</sup><https://github.com/LAL-project/linear-arrangement-library/>



Figure 4:  
Illustration of an edge's anchor  $\alpha_{ru}(\pi)$   
and coanchor  $\beta_{ru}(\pi)$ . In this figure,  
 $u, v, w \in \Gamma(r)$ . Figure adapted  
from Alemany-Puig and Ferrer-i-Cancho 2022



## 2.2

### Counting planar arrangements

It is well known that the number of unconstrained arrangements of an  $n$ -vertex tree is  $n!$ . This is true given that arrangements are simply permutations, and unconstrained arrangements are not subject to any particular constraint, thus all vertex orderings are possible. Building on the fact that projective arrangements span over contiguous intervals (Kuhlmann and Nivre 2006), Alemany-Puig and Ferrer-i-Cancho 2022 studied the expected value of the random variable  $D(T^r)$  in such arrangements by defining, as usual, a set of *segments*  $\Phi_u$  associated to each vertex  $u$ , consisting of the segments associated to the subtrees  $T_{u_1}^r, \dots, T_{u_p}^r$  and  $u$ . A *segment* of a rooted tree  $T_u^r$  is a segment within the linear ordering containing all vertices of  $T_u^r$ , an interval of length  $s_r(u)$  whose starting and ending positions are unknown until the whole tree is fully linearized; thus, a segment is a movable set of vertices within the linear ordering (Alemany-Puig and Ferrer-i-Cancho 2022). For a vertex  $u$ , the set  $\Phi_u$  is constructed from vertex  $u$ 's segment and the segments of its children  $\Gamma_r(u) = \{u_1, \dots, u_k\}$  (Figure 5). Decomposing every vertex and its segments from the root to the leaves linearizes  $T^r$  into a projective arrangement (Figure 5). This characterization led to a straightforward derivation of the number of projective arrangements of a rooted tree  $T^r$  (Table 1)

$$(6) \quad N_{\text{pr}}(T^r) = \prod_{u \in V} (d_r(u) + 1)!$$

Using the structure of segments summarized above, we present a characterization of planar arrangements of free trees which helps to devise a method to generate planar arrangements u.a.r. (Section 2.3.3) and to prove Theorem 1 (Section 2.4). To this aim, we define  $\mathbf{P}_{\text{pr}}^\circ(T^r)$  as the set of projective arrangements of a rooted tree  $T^r$  such that  $\pi(r) = 1$ , and denote its size as  $N_{\text{pr}}^\circ(T^r) = |\mathbf{P}_{\text{pr}}^\circ(T^r)|$ . Notice that



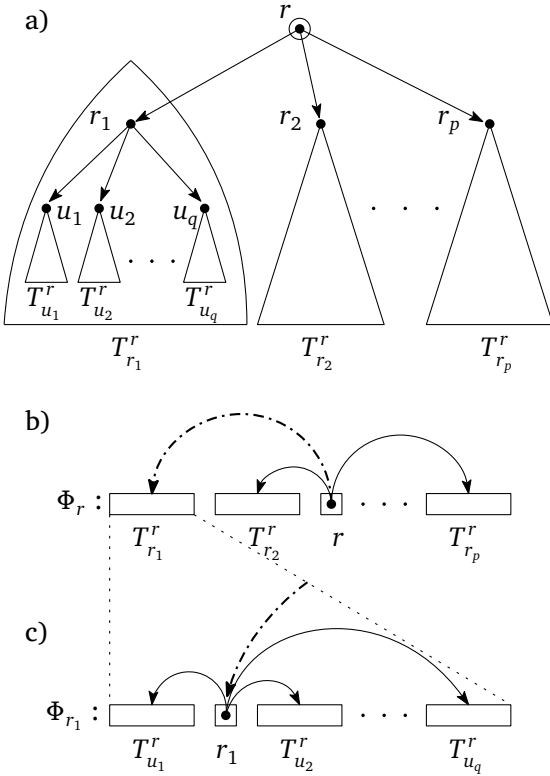


Figure 5:  
 a) A rooted tree  $T^r$  where  $\Gamma(r) = \{r_1, \dots, r_p\}$  are the  $p$  children of  $r$ . The subtree  $T_{r_1}^r$  has been circled for clarity. b) An example of a permutation of the segments in  $\Phi_r$  associated to the root. c) An example of a permutation of the segments in  $\Phi_{r_1}$  associated to  $r_1$ , the segment at the leftmost position in the example in (b). The dash-dotted edge in (b) and in (c) represent the same edge of the tree. In (b) and (c), respectively,  $r$  and  $r_1$  are segments of length 1

when a vertex  $u$  is fixed to the leftmost position, the planar arrangements in  $\mathbf{P}_{\text{pr}}^\diamond(T^u)$  are obtained by arranging the subtrees  $T_v^u, v \in \Gamma(u)$ , projectively to the right of  $u$  in the linear arrangement. It is important to bear in mind that the operator  $\diamond$  only fixes the root vertex  $r$  to the leftmost position of the arrangement: the other vertices can be placed freely as long as the result is projective.

**PROPOSITION 1** *The number of planar arrangements of an  $n$ -vertex free tree  $T = (V, E)$ , with  $V = \{u_1, \dots, u_n\}$  is*

$$(7) \quad \mathbf{N}_{\text{pl}}(T) = n\mathbf{N}_{\text{pr}}^\diamond(T^{u_1}) = \dots = n\mathbf{N}_{\text{pr}}^\diamond(T^{u_n}) = n \prod_{u \in V} d(u)!$$

**PROOF** Given a free tree  $T$ , and any two distinct vertices  $u, v$ , it holds that  $\mathbf{P}_{\text{pr}}^\diamond(T^u) \cap \mathbf{P}_{\text{pr}}^\diamond(T^v) = \emptyset$  because the vertices in the first positions are different. This lets us partition  $\mathbf{P}_{\text{pl}}(T)$  into the non-empty

pairwise-disjoint sets  $\mathbf{P}_{\text{pr}}^\circ(T^u)$  and see that

$$N_{\text{pl}}(T) = \sum_{u \in V} N_{\text{pr}}^\circ(T^u).$$

It is easy to see that

$$N_{\text{pr}}^\circ(T^u) = d(u)! \prod_{v \in \Gamma(u)} N_{\text{pr}}^\circ(T_v^u) = \prod_{v \in V} d(v)!.$$

We used Equation 6 in the second equality. Notice that

$$N_{\text{pr}}^\circ(T^{u_1}) = \dots = N_{\text{pr}}^\circ(T^{u_n}),$$

since the value  $N_{\text{pr}}^\circ(T^u)$  does not depend on the root vertex  $u$ . Therefore, Equation 7 follows immediately.  $\square$

Obviously, there are more planar arrangements of a free tree  $T$  than projective arrangements of any ‘rooting’  $T^r$  of  $T$ , formally  $N_{\text{pl}}(T) \geq N_{\text{pr}}(T^r)$ . We can see this by noticing that, when given a ‘rooting’ of  $T$  at  $r \in V$ ,

$$\frac{N_{\text{pl}}(T)}{N_{\text{pr}}(T^r)} = \frac{nd(r)! \prod_{u \in V \setminus \{r\}} d(u)!}{(d(r)+1)! \prod_{u \in V \setminus \{r\}} d(u)!} = \frac{n}{d(r)+1} \geq 1,$$

with equality when  $T$  is a star tree<sup>4</sup> and  $r$  is its vertex of highest degree.

### 2.3 *Generating arrangements uniformly at random*

Arrangements can be generated freely, that is, by imposing no constraint on the possible orderings, where all the  $n!$  possible orderings are equally likely, or by imposing some constraint on the possible orderings. Generating unconstrained arrangements is straightforward: it is well known that a permutation of  $n$  elements can be generated u.a.r. in time  $O(n)$  (Cormen *et al.* 2001). It can be done as follows. Assume we are given a set of  $n$  vertices, say  $V = \{u_1, \dots, u_n\}$ , and let  $i = 1$ . Repeat the following steps  $n$  times:

---

<sup>4</sup> An  $n$ -vertex star tree consists of a vertex connected to  $n-1$  leaves; it is also a complete bipartite graph  $K_{1,n-1}$ .

1. select u.a.r. a vertex from  $V$ ; the vertex is chosen with probability  $1/(n - i + 1)$ . Let  $u_i$  be said vertex,
2. place  $u_i$  in the arrangement at position  $i$ , that is, let  $\pi(u_i) = i$ ,
3. remove  $u_i$  from  $V$ ,
4. increment  $i$  by 1.

The product of all probabilities of vertex choice gives that the probability of producing a certain linear arrangement is

$$\prod_{i=1}^n \frac{1}{n - i + 1} = \frac{1}{n!}$$

thus the arrangement is constructed uniformly at random. Since the removal of a vertex from the set and uniformly random choice of vertex can both be implemented in constant time (using arrays), the running time is  $O(n)$ .

When constraints are involved, projectivity is often the preferred choice (Gildea and Temperley 2007; Liu 2008; Futrell *et al.* 2015). First, we present a  $O(n)$ -time procedure to generate projective arrangements u.a.r. (Section 2.3.1) and review methods used in past research (Section 2.3.2). Then we present a novel  $O(n)$ -time procedure to generate planar arrangements u.a.r. (Section 2.3.3) which in turn involves the generation of random projective arrangements of a subtree.

### Generating projective arrangements

2.3.1

The method we will present in detail here was outlined first by Futrell *et al.* 2015. Here we borrow from recent theoretical research summarized above (Alemany-Puig and Ferrer-i-Cancho 2022) to derive a detailed algorithm to generate projective arrangements and prove its correctness. In order to generate projective arrangements u.a.r., simply make random permutations of a vertex  $u$  and its children  $\Gamma_r(u)$ , that is, choose one of the possible  $(d_r(u) + 1)!$  permutations u.a.r. Algorithm 1 formalizes this brief description. The proof that Algorithm 1 produces projective arrangements of a rooted tree  $T^r$  u.a.r. is simple. The first call takes the root and its dependents and produces a uniformly random permutation with probability  $1/(d(r) + 1)!$ . Subsequent recursive calls (in Algorithm 2) produce the corresponding

Algorithm 1:  
Generating  
projective  
arrangements  
u.a.r

```

1 Function RANDOM_PROJECTIVE_ARRANGEMENT( $T^r$ ) is
   Input:  $T^r$  a rooted tree.
   Output: A projective arrangement  $\pi$  of  $T^r$  chosen u.a.r.
2    $\pi \leftarrow$  empty  $n$ -vertex arrangement
   // Algorithm 2
3   RANDOM_PROJECTIVE_ARRANGEMENT_SUBTREE( $T^r, r, 1, \pi$ )
4   return  $\pi$ 

```

Algorithm 2:  
Generating  
projective  
arrangements  
u.a.r. of a  
subtree

```

1 Function
   RANDOM_PROJECTIVE_ARRANGEMENT_SUBTREE( $T^r, u, p, \pi$ ) is
   Input:  $T^r$  a rooted tree,  $u$  any vertex of  $T^r$ ,  $p$  the starting position
   to arrange the vertices of  $T^r_u$ ,  $\pi$  partially-constructed
   without  $T^r_u$ .
   Output:  $\pi$  partially-constructed with  $T^r_u$ .
2    $\Phi_u \leftarrow$  a random permutation of  $\Gamma_r(u) \cup \{u\}$ 
3   for  $v \in \Phi_u$  do
4     if  $v = u$  then
5        $\pi(v) \leftarrow p$ 
6        $p \leftarrow p + 1$ 
7     else
8       RANDOM_PROJECTIVE_ARRANGEMENT_SUBTREE( $T^r, v, p, \pi$ )
9        $p \leftarrow p + s_r(v)$ 

```

permutations each with its respective uniform probability, hence the probability of producing a particular permutation is the product of individual probabilities. Using Equation 6, we easily obtain that the probability of producing a certain projective arrangement is

$$\prod_{u \in V} \frac{1}{(d_r(u) + 1)!} = \frac{1}{\mathbf{N}_{\text{pr}}(T^r)}.$$

### 2.3.2

#### Generation of projective arrangements in past research

Algorithm 1 is equivalent to the “fully random” method used by Futrell *et al.* 2015 as witnessed by the implementation of their code available on Github,<sup>5</sup> in particular in file cliqs/mindep.py<sup>6</sup> (function `_randlin_projective`). Notice that Futrell *et al.* 2015 outline

<sup>5</sup><https://github.com/Futrell/cliqs/tree/44bfcf2c42c848243c264722b5eccdfec0ede6a>

<sup>6</sup><https://github.com/Futrell/cliqs/blob/44bfcf2c42c848243c264722b5eccdfec0ede6a/cliqs/mindep.py>

(though vaguely) that a projective arrangement is generated randomly by “Starting at the root node of a dependency tree, collect[ing] the head word and its dependents and order[ing] them randomly”.

Futrell *et al.* 2015 present their method to generate random projective arrangements as though it were the same as that by Gildea and Temperley 2007, 2010, who introduced a method to generate random linearizations of a tree which consists of “choosing a random branching direction for each dependent of each head,<sup>7</sup> and – in the case of multiple dependents on the same side – randomly ordering them in relation to the head” (Gildea and Temperley 2010). However, Futrell *et al.* 2015 do not actually implement Gildea and Temperley’s method as witnessed by their code. Critically, Gildea and Temperley’s method does not produce uniformly random linearizations as we show with a counterexample.

Consider a star tree rooted at its hub. Let  $X$  be a random variable for the position of the root in a random projective linear arrangement ( $1 \leq X \leq n$ ). We have  $\mathbb{P}(X = x) = 1/n$  for all  $x \in [1, n]$ , therefore  $X$  follows a uniform distribution and hence  $\mathbb{E}[X] = (n + 1)/2$  and  $\mathbb{V}[X] = (n^2 - 1)/12$  (Mitzenmacher and Upfal 2017). Let  $X'$  be a random variable for the position of the root according to Gildea and Temperley’s method. It is easy to see that  $X' - 1$  follows a binomial distribution with parameters  $n - 1$  and  $1/2$ . Namely,  $\mathbb{P}(X' - 1 = x) = \binom{n-1}{x}/2^{n-1}$ . We have that  $\mathbb{E}[X'] = 1 + \mathbb{E}[X' - 1] = (n + 1)/2 = \mathbb{E}[X]$ , but  $\mathbb{V}[X'] = \mathbb{V}[X' - 1] = (n - 1)/4$ . Therefore, the variance in a truly uniformly random projective linear arrangement is  $\Theta(n^2)$  while Gildea and Temperley’s method results in  $\Theta(n)$ , a much smaller dispersion. As  $n \rightarrow \infty$ ,  $X' - 1$  converges to a Gaussian distribution.

Gildea and Temperley’s method was introduced as a random baseline for the distance between syntactically-related words in languages and has been used with that purpose (Gildea and Temperley 2007, 2010; Temperley and Gildea 2018). Interestingly, the minimum baseline, namely, the minimum sum of dependency distances, results from placing the root at the center (Shiloach 1979; Chung 1984). The example above shows that Gildea and Temperley’s baseline tends to put the root at the center of the linear arrangement

---

<sup>7</sup>That is, as explained by Temperley and Gildea 2018, “choose a random assignment of each dependent to either the left or the right of its head.”

with higher probability than the truly uniform baseline. That behavior casts doubts on the power of that random baseline to investigate dependency distance minimization in languages since it tends to place the root at the center of the sentence, as expected from an optimal placement under projectivity (Gildea and Temperley 2007; Alemany-Puig *et al.* 2021) and does it with much lower dispersion around the center than in truly uniformly random linearizations.

### 2.3.3 Generating planar arrangements

Proposition 1 leads to a method to generate planar arrangements u.a.r. for any free tree  $T$ . The method we propose is detailed in Algorithm 3.

Algorithm 3:  
Generating  
planar  
arrangements  
u.a.r.

```

1 Function RANDOM_PLANAR_ARRANGEMENT( $T$ ) is
   Input:  $T$  a free tree.
   Output: A planar arrangement  $\pi$  of  $T$  chosen u.a.r.
2    $\pi \leftarrow$  empty  $n$ -vertex arrangement
3    $u \leftarrow$  a vertex of  $T$  chosen u.a.r.
4    $\pi(u) \leftarrow 1$ 
5    $\Phi_u \leftarrow$  a random permutation of  $\Gamma(u)$ 
6    $p \leftarrow 2$ 
7   for  $v \in \Phi_u$  do
   |   // Algorithm 2
   |   RANDOM_PROJECTIVE_ARRANGEMENT_SUBTREE( $T^u, v, p, \pi$ )
   |    $p \leftarrow p + s_u(v)$ 
10  return  $\pi$ 

```

It is easy to see that Algorithm 3 has time complexity  $O(n)$ . Now we show that it generates planar arrangements uniformly at random. Firstly, choose a vertex, say  $u \in V$ , u.a.r., and place it at one of the arrangement's ends, say, the leftmost position; this vertex acts as a root for  $T$ . Secondly, choose u.a.r. one of the  $d(u)!$  permutations of the segments of the subtrees  $T_v^u$  u.a.r. Lastly, recursively choose u.a.r. a projective linearization of every subtree  $T_v^u$  for  $v \in \Gamma(u)$  (Algorithm 2). These steps generate a planar arrangement u.a.r. since the probability of producing a certain planar arrangement following these steps is, then,

$$\frac{1}{n} \frac{1}{d(u)!} \prod_{v \in \Gamma(u)} \frac{1}{\mathbf{N}_{\text{pr}}(T_v^u)} = \frac{1}{n} \frac{1}{d(u)!} \prod_{v \in V \setminus \{u\}} \frac{1}{d(v)!} = \frac{1}{\mathbf{N}_{\text{pl}}(T)}.$$

The equalities follow from Proposition 1.

In this section we derive an arithmetic expression for  $\mathbb{E}_{\text{pl}}[D(T)]$ . First, we prove Theorem 1. To this aim, we define  $\mathbb{E}_{\text{pr}}^{\circ}[\alpha_{uv} | r] = \mathbb{E}_{\text{pr}}[\alpha_{uv} | \pi(r) = 1]$  as the expected value of  $\alpha_{uv}$  conditioned to the projective arrangements  $\pi$  of  $T^r$  such that  $\pi(r) = 1$ ; we define  $\mathbb{E}_{\text{pr}}^{\circ}[\beta_{uv} | r]$  likewise. The root is specified as a parameter of the expected value because we want to be able to use various roots. In the following proofs we rely heavily on Linearity of Expectation (Mitzenmacher and Upfal 2017, Theorem 2.1) and the Law of Total Expectation (Mitzenmacher and Upfal 2017, Lemma 2.5).

**PROOF** [Proof of Theorem 1] We first prove Equation 4. By the Law of Total Expectation,

$$\mathbb{E}_{\text{pl}}[D(T)] = \sum_{u \in V} \mathbb{E}_{\text{pl}}[D(T) | \pi(u) = 1] \mathbb{P}_{\text{pl}}(\pi(u) = 1).$$

Notice, quite simply, that

$$\mathbb{E}_{\text{pl}}[D(T) | \pi(u) = 1] = \mathbb{E}_{\text{pr}}[D(T^u) | \pi(u) = 1] = \mathbb{E}_{\text{pr}}^{\circ}[D(T^u)],$$

that is, the expected value of  $D$  conditioned to planar arrangements of  $T$  such that  $u$  is fixed at the leftmost position,  $\mathbb{E}_{\text{pl}}[D(T) | \pi(u) = 1]$ , is equal to the expected value of  $D$  conditioned to projective arrangements of  $T^u$  such that vertex  $u$  is fixed at the leftmost position, which is denoted as  $\mathbb{E}_{\text{pr}}^{\circ}[D(T^u)]$ . By noticing, given a fixed vertex  $u$ , that  $\mathbb{P}_{\text{pl}}(\pi(u) = 1) = \frac{1}{n}$ , which is the proportion of planar arrangements of  $T$  in which  $\pi(u) = 1$  (Proposition 1), Equation 4 follows immediately. Notice that Equation 4 expresses the expected value of  $D$  conditioned to planar arrangements of a free tree  $T$  as the average of each of the expected values of  $D$  conditioned to projective arrangements of  $T^u$  (for all  $u \in V$ ) such that the root is fixed at the leftmost position.

Now we aim to write  $\mathbb{E}_{\text{pr}}^{\circ}[D(T^u)]$  as a function of  $\mathbb{E}_{\text{pr}}[D(T^u)]$ . We start by decomposing  $\mathbb{E}_{\text{pr}}^{\circ}[D(T^u)]$  into a summation of expected values of the individual edge lengths, and group the edges of every subtree  $T_v^u$  of  $T^u$  (where  $uv$  is a (directed) edge of the tree) into one single expected value for each subtree and leave the edges incident to the root  $u$  in the same summation as follows:

$$\mathbb{E}_{\text{pr}}^{\circ}[D(T^u)] = \sum_{vw \in \Gamma(u)} \left( \mathbb{E}_{\text{pr}}^{\circ}[\delta_{vw} | u] + \mathbb{E}_{\text{pr}}[D(T_v^u)] \right).$$

Now, it is important to notice that we did not write  $\mathbb{E}_{\text{pr}}^{\diamond}[D(T_v^u)]$  in the summation above since the conditioning imposed by the operator  $\diamond$  in  $\mathbb{E}_{\text{pr}}^{\diamond}[D(T^u)]$  only applies to the root  $u$ . The root of the subtrees can be placed freely in the arrangement as long as the result is projective. Now we decompose all (directed) edges  $uv$  of  $T^r$  in the first summation into anchor and coanchor, and we get

$$\mathbb{E}_{\text{pr}}^{\diamond}[D(T^u)] = \sum_{v \in \Gamma(u)} \left( \mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} + \beta_{uv} \mid u] + \mathbb{E}_{\text{pr}}^{\diamond}[D(T_v^u)] \right).$$

Although the root  $u$  is clear in this context, we have made it explicit in  $\mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} + \beta_{uv} \mid u]$  so as to be able to keep track of it in the following derivations. By linearity of expectation,

$$\mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} + \beta_{uv} \mid u] = \mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} \mid u] + \mathbb{E}_{\text{pr}}^{\diamond}[\beta_{uv} \mid u].$$

Now, notice that the length of the anchor of any given directed edge  $(u, v)$ , where  $u$  is the head and  $v$  is the dependent, is invariant to the position of  $u$ , that is, it only changes if we change the position of  $v$  within its interval. Therefore, fixing the head to the leftmost position of the arrangement (or any position outside the segment of  $v$ ) does not affect the value of  $\mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} \mid u]$  and we simply have that  $\mathbb{E}_{\text{pr}}^{\diamond}[\alpha_{uv} \mid u] = \mathbb{E}_{\text{pr}}[\alpha_{uv} \mid u]$  and thus

$$\begin{aligned} \mathbb{E}_{\text{pr}}^{\diamond}[D(T^u)] &= \sum_{v \in \Gamma(u)} \left( \mathbb{E}_{\text{pr}}[\alpha_{uv} \mid u] + \mathbb{E}_{\text{pr}}^{\diamond}[\beta_{uv} \mid u] \right. \\ &\quad \left. + \mathbb{E}_{\text{pr}}[D(T_v^u)] \right). \end{aligned}$$

The next step is to find the value of  $\mathbb{E}_{\text{pr}}^{\diamond}[\beta_{uv} \mid u]$ . Notice now that the length of the coanchor of any directed edge  $(u, v)$  is affected by the position of the head  $u$  and, as such,  $\mathbb{E}_{\text{pr}}^{\diamond}[\beta_{uv} \mid u]$  need not be exactly equal to  $\mathbb{E}_{\text{pr}}[\beta_{uv} \mid u]$ . The derivation is found in Appendix 4.3 since it is merely an adaptation of the proof by Alemany-Puig and Ferrer-i-Cancho 2022, Lemma 1; it gives

$$\mathbb{E}_{\text{pr}}^{\diamond}[\beta_{uv} \mid u] = \frac{3}{2} \mathbb{E}_{\text{pr}}[\beta_{uv} \mid u].$$



Thus,

$$\begin{aligned}
 \mathbb{E}_{\text{pr}}^{\circ}[D(T^u)] &= \sum_{v \in \Gamma(u)} \left( \mathbb{E}_{\text{pr}}[\alpha_{uv} | u] + \frac{3}{2} \mathbb{E}_{\text{pr}}[\beta_{uv} | u] \right. \\
 &\quad \left. + \mathbb{E}_{\text{pr}}[D(T_v^u)] \right) \\
 &= \sum_{v \in \Gamma(u)} \left( \mathbb{E}_{\text{pr}}[\delta_{uv} | u] + \mathbb{E}_{\text{pr}}[D(T_v^u)] \right) \\
 &\quad + \frac{1}{2} \mathbb{E}_{\text{pr}}[\beta_{uv} | u] \\
 (8) \qquad &= \mathbb{E}_{\text{pr}}[D(T^u)] + \frac{1}{2} \sum_{v \in \Gamma(u)} \mathbb{E}_{\text{pr}}[\beta_{uv} | u].
 \end{aligned}$$

In the third equality we have used the identity by Alemany-Puig and Ferrer-i-Cancho 2022, Equation 28, which states that in a rooted tree  $T^r$

$$\mathbb{E}_{\text{pr}}[D(T^r)] = \sum_{v \in \Gamma(r)} \left( \mathbb{E}_{\text{pr}}[\delta_{rv}] + \mathbb{E}_{\text{pr}}[D(T_v^r)] \right).$$

In this equation, we have not specified the expected values as being conditioned by the root  $r$  since this is clear from the context. Plugging Equation 8 into Equation 4 we get

$$(9) \quad \mathbb{E}_{\text{pl}}[D(T)] = \frac{1}{2n} \sum_{u \in V} \sum_{v \in \Gamma(u)} \mathbb{E}_{\text{pr}}[\beta_{uv} | u] + \frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}}[D(T^u)].$$

We can use the following result by Alemany-Puig and Ferrer-i-Cancho 2022, Equation 16:

$$\mathbb{E}_{\text{pr}}[\beta_{uv} | u] = \frac{s_u(u) - s_u(v) - 1}{3} = \frac{n - s_u(v) - 1}{3}$$

to further simplify Equation 9 and, after proving that

$$\begin{aligned}
 \sum_{v \in \Gamma(u)} \mathbb{E}_{\text{pr}}[\beta_{uv} | u] &= \sum_{v \in \Gamma(u)} \frac{s_u(u) - s_u(v) - 1}{3} \\
 &= \frac{(n-1)(d(u)-1)}{3}, \\
 \sum_{u \in V} \frac{1}{3} (n-1)(d(u)-1) &= \frac{(n-1)(n-2)}{3},
 \end{aligned}$$

we obtain

$$(10) \quad \frac{1}{2n} \sum_{u \in V} \sum_{v \in \Gamma(u)} \mathbb{E}_{\text{pr}}[\beta_{uv} | u] = \frac{(n-1)(n-2)}{6n}.$$

Hence Equation 5. □

For the sake of comprehensiveness, we also provide an arithmetic expression for the expected length of an edge  $uv$  of a free tree in uniformly random planar arrangements. To this aim, we further define  $\mathbb{E}_{\text{pl}}^\circ[\delta_{uv} | r] = \mathbb{E}_{\text{pl}}[\delta_{uv} | \pi(r) = 1]$  to be the expected value of the length of edge  $uv \in E(T)$  when the vertex  $r \in V(T)$  is fixed to the leftmost position in planar arrangements of  $T$ . Similarly, given a rooting of  $T$  at  $r$ , let  $\mathbb{E}_{\text{pr}}^\circ[\delta_{uv} | r] = \mathbb{E}_{\text{pr}}[\delta_{uv} | \pi(r) = 1]$  to be the expected value of the length of edge  $uv \in E(T^r)$  when vertex  $r$  acts as the root of the tree and it is fixed to the leftmost position in projective arrangements of  $T^r$ . The root vertex  $r$  may be vertex  $u$ , vertex  $v$ , or neither. In the expected value  $\mathbb{E}_{\text{pr}}^\circ[\delta_{uv} | r]$  we assume that the edge  $uv$  is directed from  $u$  to  $v$  in accordance with the orientation defined by the root vertex  $r$ . Therefore, when  $r$  is neither  $u$  nor  $v$ , the vertex of edge  $uv$  closest to  $r$  is always vertex  $u$ , and the farthest is always vertex  $v$ .

**LEMMA 2**     *Given a free tree  $T = (V, E)$ , for any  $uv \in E$  it holds that*

$$(11) \quad \mathbb{E}_{\text{pl}}[\delta_{uv}] = 1 + \frac{1}{n} \sum_{r \in V \setminus \{u, v\}} \mathbb{E}_{\text{pr}}[\delta_{uv} | r],$$

where as per Alemany-Puig and Ferrer-i-Cancho 2022

$$(12) \quad \mathbb{E}_{\text{pr}}[\delta_{uv} | r] = \frac{2s_r(u) + s_r(v) + 1}{6}.$$

**PROOF**     Following the characterization of planar arrangements described in Section 2.2, we have that  $\mathbb{P}_{\text{pl}}(\pi(r) = 1) = 1/n$ . Then applying the Law of Total Expectation

$$(13) \quad \begin{aligned} \mathbb{E}_{\text{pl}}[\delta_{uv}] &= \sum_{r \in V} \mathbb{E}_{\text{pl}}[\delta_{uv} | \pi(r) = 1] \mathbb{P}_{\text{pl}}(\pi(r) = 1) \\ &= \frac{1}{n} \sum_{r \in V} \mathbb{E}_{\text{pl}}^\circ[\delta_{uv} | r]. \end{aligned}$$

Now we calculate  $\mathbb{E}_{\text{pl}}^\circ[\delta_{uv} | r]$  by cases. When  $r \notin \{u, v\}$ ,

$$(14) \quad \mathbb{E}_{\text{pl}}^\circ[\delta_{uv} | r] = \mathbb{E}_{\text{pr}}^\circ[\delta_{uv} | r] = \mathbb{E}_{\text{pr}}[\delta_{uv} | r].$$

When  $r \in \{u, v\}$ , by linearity of expectation,

$$\begin{aligned} \mathbb{E}_{\text{pl}}^\circ[\delta_{uv} \mid r] &= \mathbb{E}_{\text{pr}}^\circ[\delta_{uv} \mid r] \\ &= \mathbb{E}_{\text{pr}}^\circ[\alpha_{uv} + \beta_{uv} \mid r] \\ &= \mathbb{E}_{\text{pr}}^\circ[\alpha_{uv} \mid r] + \mathbb{E}_{\text{pr}}^\circ[\beta_{uv} \mid r]. \end{aligned}$$

By denoting  $\bar{r}$  the only vertex in  $\{u, v\} \setminus \{r\}$ , then

$$(15) \quad \mathbb{E}_{\text{pr}}^\circ[\alpha_{uv} \mid r] = \mathbb{E}_{\text{pr}}[\alpha_{uv} \mid r] = \frac{s_r(\bar{r}) + 1}{2}.$$

Equation 15 relies on the fact that in a rooted tree  $T^r$ , the expected length of the anchor of an edge incident to the root, say  $rw \in E(T^r)$ , is given by  $\mathbb{E}_{\text{pr}}[\alpha_{rw} \mid r] = (s_r(w) + 1)/2$  (Alemany-Puig and Ferrer-i-Cancho 2022). An arithmetic expression for  $\mathbb{E}_{\text{pr}}^\circ[\beta_{uv} \mid r]$  can be found by modifying the proof of Alemany-Puig and Ferrer-i-Cancho 2022, Lemma 1. Then, as before, we get (see Appendix 4.3),

$$(16) \quad \mathbb{E}_{\text{pr}}^\circ[\beta_{uv} \mid r] = \frac{3}{2} \mathbb{E}_{\text{pr}}[\beta_{uv} \mid r] = \frac{n - s_r(\bar{r}) - 1}{2}.$$

Therefore, by adding Equations 15 and 16 we obtain

$$\begin{aligned} \mathbb{E}_{\text{pl}}^\circ[\delta_{uv} \mid r] &= \mathbb{E}_{\text{pr}}^\circ[\alpha_{uv} \mid r] + \mathbb{E}_{\text{pr}}^\circ[\beta_{uv} \mid r] \\ &= \frac{s_r(\bar{r}) + 1}{2} + \frac{n - s_r(\bar{r}) - 1}{2} \\ (17) \quad &= \frac{n}{2}. \end{aligned}$$

Equation 11 follows immediately after inserting Equations 17 and 14 in Equation 13.  $\square$

## APPLICATIONS

3

### *A linear-time algorithm to compute $\mathbb{E}_{\text{pl}}[D(T)]$*

3.1

Here we consider algorithms of increasing efficiency. First, since  $\mathbb{E}_{\text{pr}}[D(T^u)]$  can be calculated in  $O(n)$ -time for any  $n$ -vertex rooted tree  $T^u$  (Alemany-Puig and Ferrer-i-Cancho 2022, Theorem 1), the evaluation ‘as is’ of Equation 5 leads to a  $O(n^2)$ -time algorithm.

Second, we could calculate the value  $\mathbb{E}_{\text{pr}}[D(T^u)]$  for all  $u \in V$  in  $O(n)$ -time and  $O(n)$ -space with the following procedure:

1. Precompute  $s_u(v)$  in  $O(n)$ -time (Alemany-Puig *et al.* 2022);
2. Choose an arbitrary vertex  $w$ ;
3. Calculate  $\mathbb{E}_{\text{pr}}[D(T^w)]$  in  $O(n)$ -time (Alemany-Puig and Ferrer-i-Cancho 2022); and, finally,
4. Perform a Breadth First Search (BFS) traversal of  $T$  starting at  $w$ . In this traversal, when going from vertex  $u$  to vertex  $v$ , the value of  $\mathbb{E}_{\text{pr}}[D(T^v)]$  is calculated applying the precomputed value of  $\mathbb{E}_{\text{pr}}[D(T^u)]$  to the following equation:

$$\mathbb{E}_{\text{pr}}[D(T^u)] = \mathbb{E}_{\text{pr}}[D(T^v)] + \Delta,$$

where  $\Delta$  is equal to the difference  $\mathbb{E}_{\text{pr}}[D(T^u)] - \mathbb{E}_{\text{pr}}[D(T^v)]$ . We can obtain a formula for this difference by manipulating Equation 3. We get

$$\begin{aligned} \Delta &= \mathbb{E}_{\text{pr}}[D(T^u)] - \mathbb{E}_{\text{pr}}[D(T^v)] \\ &= \frac{1}{6} [s_u(v)(2d(v) - 1) + 2n(d(u) - d(v)) \\ &\quad - s_v(u)(2d(u) - 1)]. \end{aligned}$$

Notice that the value of  $\Delta$  can be computed in constant time for any two vertices  $u$  and  $v$  (here we are interested in the value of  $\Delta$  for pairs of adjacent vertices) and, crucially, without knowledge of either  $\mathbb{E}_{\text{pr}}[D(T^u)]$  or  $\mathbb{E}_{\text{pr}}[D(T^v)]$ . That is, if the value of  $\mathbb{E}_{\text{pr}}[D(T^u)]$  is known then the value of  $\mathbb{E}_{\text{pr}}[D(T^v)]$  for any  $v \in \Gamma(u)$  can be calculated in constant time as

$$\mathbb{E}_{\text{pr}}[D(T^v)] = \mathbb{E}_{\text{pr}}[D(T^u)] - \Delta.$$

Third, we propose an alternative that is also  $O(n)$ -time yet simpler and faster in practice, based on Proposition 2.

**PROPOSITION 2**     *Given a free tree  $T = (V, E)$ ,*

$$(18) \quad \mathbb{E}_{\text{pl}}[D(T)] = \frac{(n-1)(3n^2 + 2n - 2)}{6n} - \frac{1}{6n} \sum_{v \in V} (2d(v) - 1) \sum_{u \in \Gamma(v)} s_v(u)^2.$$

**PROOF** Here we simplify the summation in Equation 5, which becomes (as per Alemany-Puig and Ferrer-i-Cancho 2022)

$$\frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}} [D(T^u)] = \frac{1}{6n} (f(T) - n)$$

with

$$f(T) = \sum_{u \in V} \sum_{v \in V} s_u(v)(d_u(v) + 1).$$

Now we simplify  $f(T)$  by first replacing the term  $d_u(v)$  by  $d(v)$  after the necessary transformations so that we can swap the order of the summations afterwards, that is,

$$\begin{aligned} f(T) &= \sum_{u \in V} (s_u(u)(2d_u(u) + 1) + \sum_{v \in V \setminus \{u\}} s_u(v)(2d_u(v) + 1)) \\ &= \sum_{u \in V} n(2d(u) + 1) + \sum_{u \in V} \sum_{v \in V \setminus \{u\}} s_u(v)(2d(v) - 1) \\ &= n(5n - 4) - \sum_{u \in V} s_u(u)(2d(u) - 1) \\ &\quad + 2 \sum_{u \in V} \sum_{v \in V} s_u(v)d(v) - \sum_{u \in V} \sum_{v \in V} s_u(v) \\ (19) \quad &= 2n^2 + g(T) - h(T) \end{aligned}$$

with

$$(20) \quad g(T) = 2 \sum_{u \in V} \sum_{v \in V} s_u(v)d(v),$$

$$(21) \quad h(T) = \sum_{u \in V} \sum_{v \in V} s_u(v).$$

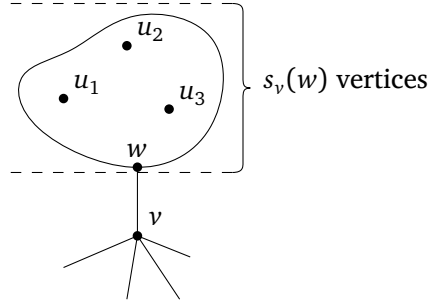
In the preceding derivation, the second equality holds due to  $d_u(v) = d(v) - 1$  for  $v \neq u$ ; the third and fourth steps, we apply the Handshaking lemma.<sup>8</sup> These lead to

$$(22) \quad \frac{1}{n} \sum_{u \in V} \mathbb{E}_{\text{pr}} [D(T^u)] = \frac{1}{6n} (n(2n - 1) + g(T) - h(T)).$$

---

<sup>8</sup>The Handshaking lemma (Gunderson 2014) states that the sum of the degrees of all vertices of a graph equals twice the number of its edges.

Figure 6:  
 Proof of Proposition 2. The value  $s_u(v)$  is the same for all vertices of  $T_w^v$  denoted as  $\{u_1, \dots, u_k\}$  in the figure and the proof



It remains to simplify Equations 20 and 21. We start by changing the order of the summations in Equation 20,

$$g(T) = 2 \sum_{v \in V} \sum_{u \in V} s_u(v) d(v) = 2 \sum_{v \in V} d(v) \sum_{u \in V} s_u(v),$$

and continue simplifying the inner summation. Consider a fixed  $v \in V$ . We have that

$$\underbrace{\sum_{u \in V} s_u(v)}_{(1)} = n + \underbrace{\sum_{u \in V \setminus \{v\}} s_u(v)}_{(2)} = n + \sum_{w \in \Gamma(v)} s_w(v) s_v(w).$$

The summation (1) adds up the size of all subtrees  $T_w^v$  with respect to a ‘moving’ root  $w$ . In the first equality we have simply taken out the case  $s_u(u)$ . To understand the second equality, focus for now on a single subtree  $T_w^v$  such that  $wv \in E$ . The summation (2) contains summands that correspond to all the vertices in  $T_w^v$ , say vertices  $u_1, \dots, u_k$  (assume, without loss of generality, that  $w = u_k$ ). These summands are  $s_{u_1}(v), \dots, s_{u_k}(v)$ , which are all equal to  $s_w(v)$  (Figure 6). Moreover, there are  $s_v(w)$  vertices in  $T_w^v$  thus  $k = s_v(w)$ , and this holds for all  $w \in \Gamma(v)$ , hence the equality. Finally,

$$(23) \quad \sum_{u \in V} s_u(v) = n + \sum_{u \in \Gamma(v)} (n - s_v(u)) s_v(u) = n^2 - \sum_{u \in \Gamma(v)} s_v(u)^2,$$

thanks to the identity  $s_u(v) + s_v(u) = n$ . Then,

$$(24) \quad g(T) = 4n^2(n-1) - 2 \sum_{v \in V} d(v) \sum_{u \in \Gamma(v)} s_v(u)^2.$$

We use the result in Equation 23 to simplify Equation 21,

$$(25) \quad h(T) = \sum_{v \in V} \sum_{u \in V} s_u(v) = n^3 - \sum_{v \in V} \sum_{u \in \Gamma(v)} s_v(u)^2.$$

By combining Equations 24 and 25 into Equation 22 and, after some effort, we obtain

$$\mathbb{E}_{\text{pl}}[D(T)] = \frac{(n-1)(n-2)}{6n} + \frac{1}{6n} \left( n(n-1)(3n+1) - \sum_{v \in V} (2d(v)-1) \sum_{u \in \Gamma(v)} s_v(u)^2 \right),$$

which leads directly to Equation 18.  $\square$

**LEMMA 3** For any given free tree  $T$ , Algorithm 4 calculates  $\mathbb{E}_{\text{pl}}[D(T)]$  in time and space  $O(n)$ .

**PROOF** The pseudocode to calculate  $\mathbb{E}_{\text{pl}}[D(T)]$  based on Proposition 2 is given in Algorithm 4. This algorithm first calculates  $s_u(v)$  for all edges  $uv \in E$ , for the given tree  $T$  in  $O(n)$  time using the pseudocode by Alemany-Puig *et al.* 2022, Algorithm 2.1. Then it uses these values to calculate the sums of  $s_v(u)^2$  for every vertex  $v \in V$ . Such sums are then used to evaluate Equation 18 hence calculating  $\mathbb{E}_{\text{pl}}[D(T)]$  in time  $O(n)$ .  $\square$

```

1 Function COMPUTE_EXPECTED_PLANAR( $T$ ) is
   Input:  $T$  free tree.
   Output:  $\mathbb{E}_{\text{pl}}[D(T)]$ .
   // Alemany-Puig et al. 2022, Algorithm 2.1
2    $S \leftarrow$  COMPUTE_S_FT( $T$ )
3    $L \leftarrow \{0\}^n$  // a vector of  $n$  zeroes.
4   for  $(u, v, s_u(v)) \in S$  do  $L[u] \leftarrow L[u] + s_u(v)^2$ 
5   return  $((n-1)(3n^2 + 2n - 2) - \sum_{u \in V} (d(u) - 1)L[u]) / 6n$ 

```

Algorithm 4:  
Calculation  
of  $\mathbb{E}_{\text{pl}}[D(T)]$ .  
Cost  $O(n)$ -time,  
 $O(n)$ -space

### A simple application

### 3.1.1

Let  $\mathbb{E}_{\geq 1}[D(T)]$  be the expected value of the sum of edge lengths conditioned to arrangements  $\pi$  such that  $C_\pi(T) \geq 1$ . That is, arrangements such that the number of edge crossings is at least 1. An immediate consequence of Lemma 3 is that  $\mathbb{E}_{\geq 1}[D(T)]$  can be computed easily as the following corollary states.

**COROLLARY 3** For any free tree  $T$ ,  $\mathbb{E}_{\geq 1}[D(T)]$  can be computed in time and space  $O(n)$  thanks to the fact that

$$(26) \quad \mathbb{E}_{\geq 1}[D(T)] = \frac{\mathbb{E}[D(T)] - \mathbb{E}_{\text{pl}}[D(T)]\mathbb{P}(C(T) = 0)}{\mathbb{P}(C(T) \geq 1)}$$

with  $\mathbb{P}(C(T) \leq 0) = \mathbf{N}_{\text{pl}}(T)/n!$  and  $\mathbb{P}(C(T) \geq 1) = (n! - \mathbf{N}_{\text{pl}}(T))/n!$ .

**PROOF** Due to the Law of Total Expectation,

$$(27) \quad \mathbb{E}[D(T)] = \mathbb{E}_{\text{pl}}[D(T)]\mathbb{P}(C(T) = 0) + \mathbb{E}_{\geq 1}[D(T)]\mathbb{P}(C(T) \geq 1),$$

and hence Equation 26.  $\mathbf{N}_{\text{pl}}(T)$  can be computed in  $O(n)$ -time with Equation 6 and  $\mathbb{E}_{\text{pl}}[D(T)]$  can be computed in time and space  $O(n)$  (Lemma 3). Hence all the components in the right hand side of Equation 26 can be computed in time and space  $O(n)$ .  $\square$

### 3.2 Real syntactic dependency distances versus random baselines

Evidence that dependency distances are smaller than expected by chance can be obtained by random baselines of varying strength:

- None,  $\mathbb{E}[D(T)]$ , the expectation of  $D(T)$  in unconstrained random linear arrangements (Ferrer-i-Cancho 2004).
- Planarity,  $\mathbb{E}_{\text{pl}}[D(T)]$ , the expectation of  $D(T)$  in planar random linear arrangements (this article).
- Projectivity,  $\mathbb{E}_{\text{pr}}[D(T^r)]$ , the expectation of  $D(T)$  in projective random linear arrangements (Alemany-Puig and Ferrer-i-Cancho 2022; Gildea and Temperley 2007).

This raises the questions of what would be the most appropriate baseline for research on dependency distance minimization.  $\mathbb{E}_{\text{pr}}[D(T^r)]$  is by far the most widely used random baseline (Gildea and Temperley 2007; Liu 2008; Park and Levy 2009; Futrell *et al.* 2015).

Since planarity is a weaker condition than projectivity,  $\mathbb{E}_{\text{pl}}[D(T)]$  implies a gain in coverage. Accordingly, there are more planar sentences than projective sentences in real texts (Havelka 2007; Gómez-Rodríguez and g 2010, Table 1) and also in artificially-generated syntactic dependency structures (Gómez-Rodríguez *et al.* 2022, Figure 2).



However, surprisingly,  $\mathbb{E}_{\text{pl}}[D(T)]$  has never been used in research on the principle of dependency distance minimization. Here we aim to test the hypothesis that formal constraints mask the effects of the principle, a hypothesis that has already been confirmed on artificially-generated syntactic dependency structures (Gómez-Rodríguez *et al.* 2022).

Given the natural growth of dependency distance as sentence length increases (Ferrer-i-Cancho and Liu 2014; Ferrer-i-Cancho *et al.* 2022), we measure, for each sentence, the average dependency distance, namely  $\langle d \rangle = D(T)/(n - 1)$  instead of the raw total sum  $D(T)$  (a sentence of  $n$  vertices has  $n - 1$  syntactic dependencies when the structure is a tree). As, in addition to such a growth, the manifestation of the principle also depends on sentence length (the statistical bias towards shorter distances may disappear or become a bias in the opposite direction in short sentences; Ferrer-i-Cancho and Gómez-Rodríguez 2021; Ferrer-i-Cancho *et al.* 2022), we compare the actual dependency distances against the values predicted by the baselines in sentences of the same length.

#### Data and methods

#### 3.2.1

We use the Parallel Universal Dependencies 2.6 collection (Zeman *et al.* 2020) for experimentation. To control for annotation style, we consider two versions of the collection: the collection with its original content-head annotation (PUD) and its transformation into Surface-Syntactic Universal Dependencies 2.6 (hereafter PSUD). By doing so, we cover two major competing annotation styles (Gerdes *et al.* 2018).

We borrow the preprocessing methods from previous research (Ferrer-i-Cancho *et al.* 2022). The main features of the processing are that nodes that are punctuation marks are removed and that the corpus remains fully parallel after the removal (Ferrer-i-Cancho *et al.* 2022). The preprocessed data is freely available as ancillary materials of the Linear Arrangement Library website.<sup>9</sup>

With respect to previous accounts (Havelka 2007; Ferrer-i-Cancho *et al.* 2018; Gómez-Rodríguez and g 2010), our collections exhibit some remarkable statistical differences. First, the proportion of projective and planar sentences is higher in PUD, where the proportion of

<sup>9</sup><https://cqlab.upc.edu/lal/universal-dependencies/>

Table 2:  
Proportion (%)  
of projective and  
planar sentences  
in the PUD  
collection

Language	Projective	Planar	Language	Projective	Planar
Arabic	96.2	96.3	Italian	99.3	99.3
Czech	89.6	89.8	Japanese	99.7	99.7
Chinese	99.4	99.4	Korean	93.6	95.2
German	86.3	86.7	Polish	94.8	95.3
English	95.5	95.9	Portuguese	96.7	96.8
Finnish	96.4	96.7	Russian	97.6	98.0
French	98.3	98.3	Spanish	95.5	95.7
Hindi	74.3	76.3	Swedish	96.5	96.9
Icelandic	96.2	96.9	Thai	97.2	97.2
Indonesian	98.7	99.0	Turkish	93.5	94.1

Table 3:  
Proportion (%)  
of projective and  
planar sentences  
in the PSUD  
collection

Language	Projective	Planar	Language	Projective	Planar
Arabic	83.6	83.9	Italian	94.5	94.6
Czech	86.6	87.2	Japanese	35.8	35.8
Chinese	42.0	46.1	Korean	75.8	77.1
German	72.3	72.7	Polish	88.2	89.7
English	93.6	94.1	Portuguese	87.3	87.7
Finnish	88.8	89.4	Russian	95.1	95.5
French	90.5	90.6	Spanish	80.2	80.9
Hindi	43.6	44.3	Swedish	93.0	93.7
Icelandic	90.7	92.0	Thai	85.6	86.8
Indonesian	90.5	91.8	Turkish	87.6	88.3

non-projective or non-planar sentences does not exceed 10% in most cases (Tables 2 and 3). This proportion increases in PSUD; wherein, in two exceptional languages, Chinese and Hindi, it becomes larger than 50% (Table 3). Second, the difference between the proportion of non-projective and non-planar sentences is smaller than in previous reports (Gómez-Rodríguez and g 2010; Havelka 2007). Having said that, notice that our collections are fully parallel, and special care has been taken to keep annotation consistent across languages.

Given formal constraint ‘ $c$ ’ (either ‘none’, ‘planarity’ ( $c = pl$ ) or ‘projectivity’ ( $c = pr$ )) and sentence length  $n$ ,

1. We calculate  $D(T^r)$  for each  $T^r$  and also calculate the expected

sum of edge lengths under ‘ $c$ ’ different constraints (none, Equation 2; planarity, Equation 5; projectivity, Equation 3).

2. Then, for each sentence, we divide each by  $n - 1$ , to produce the mean length of its dependencies

$$\langle d_c \rangle = \frac{D}{n - 1}$$

and the expected mean of length of its dependencies under some constraint ‘ $c$ ’

$$\mathbb{E}[\langle d_c \rangle] = \frac{\mathbb{E}_c[D]}{n - 1}.$$

3. Finally, we compute the average  $\langle d_c \rangle$  and the average  $\mathbb{E}[\langle d_c \rangle]$  over all sentences of length  $n$  satisfying constraint ‘ $c$ ’.

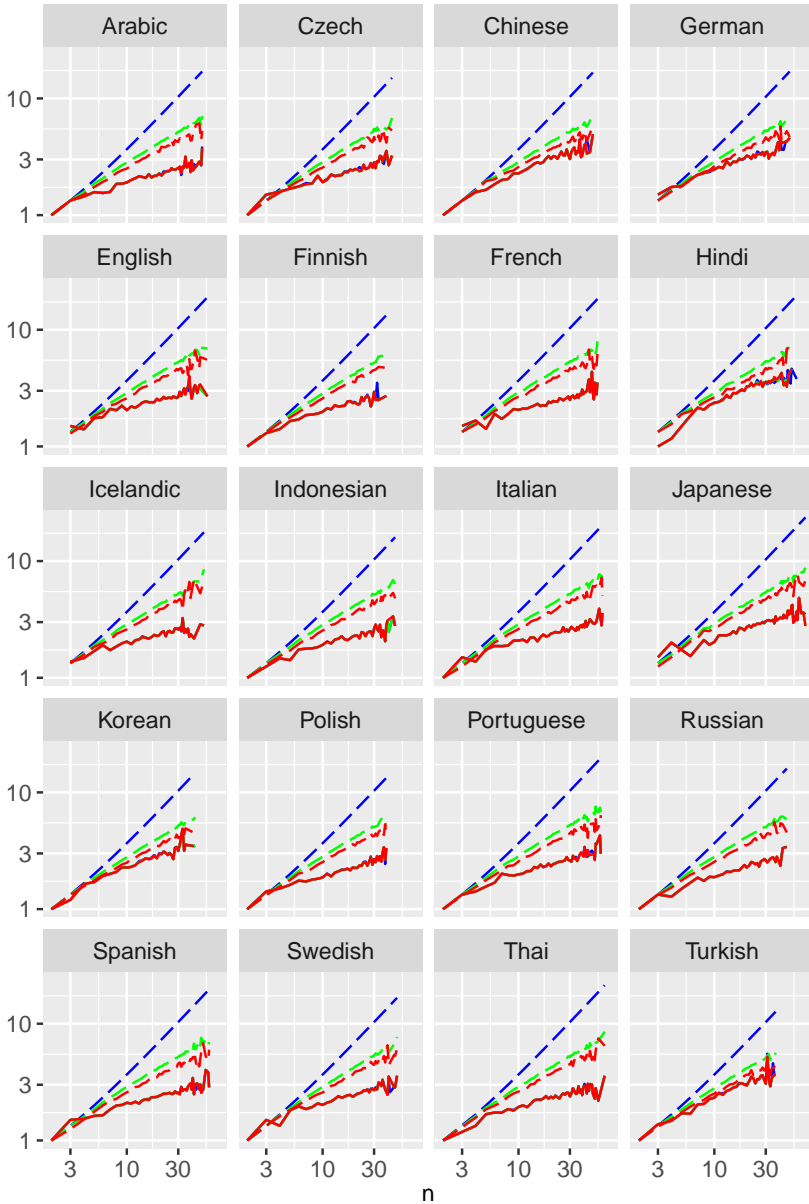
## Results

### 3.2.2

Figures 7 and 8 show the scaling of mean dependency distance as a function of sentence length in real sentences and in their corresponding random baselines. Concerning the random baselines (dashed lines), we find that the stronger the formal constraint on syntactic dependency structures, the lower the value of the random baseline. In contrast, the actual mean sentence length (solid lines) is practically the same independently of the formal constraint (none, planarity and projectivity). This is due to the fact the proportion of sentences that are lost by imposing some formal constraint is small in the PUD and PSUD collections, namely, the baselines  $\langle d \rangle$ ,  $\langle d_{pl} \rangle$  and  $\langle d_{pr} \rangle$  are extremely similar in value. The overwhelming majority of sentences are planar and the proportion of planar sentences that are not projective is really small (Table 2 and 3). Thus, selecting sentences satisfying a certain formal constraint has a negligible impact on the estimation of mean dependency distance.

Concerning the relationship between the actual mean dependency distance and the random baselines, we find that the average  $\langle d \rangle$  is below the average value of the random baselines for sufficiently large  $n$  in all languages. The only exception is Turkish, where the actual average  $\langle d \rangle$  is just slightly below the average of the projective baseline (Figures 7 and 8).

Figure 7:  
 The scaling of  $\langle d \rangle$ , the mean dependency distance of a sentence as a function of sentence length ( $n$ ) for languages in the PUD collection for formal constraints of increasing strength: none (blue), planarity (green) and projectivity (red). Lines indicate the average value over all sentences of the same length. Solid lines are used for real sentences and dashed lines are used for the corresponding random baseline. Solid lines overlap so much that only one of them can be seen in most cases



Edge lengths in random planar linearizations

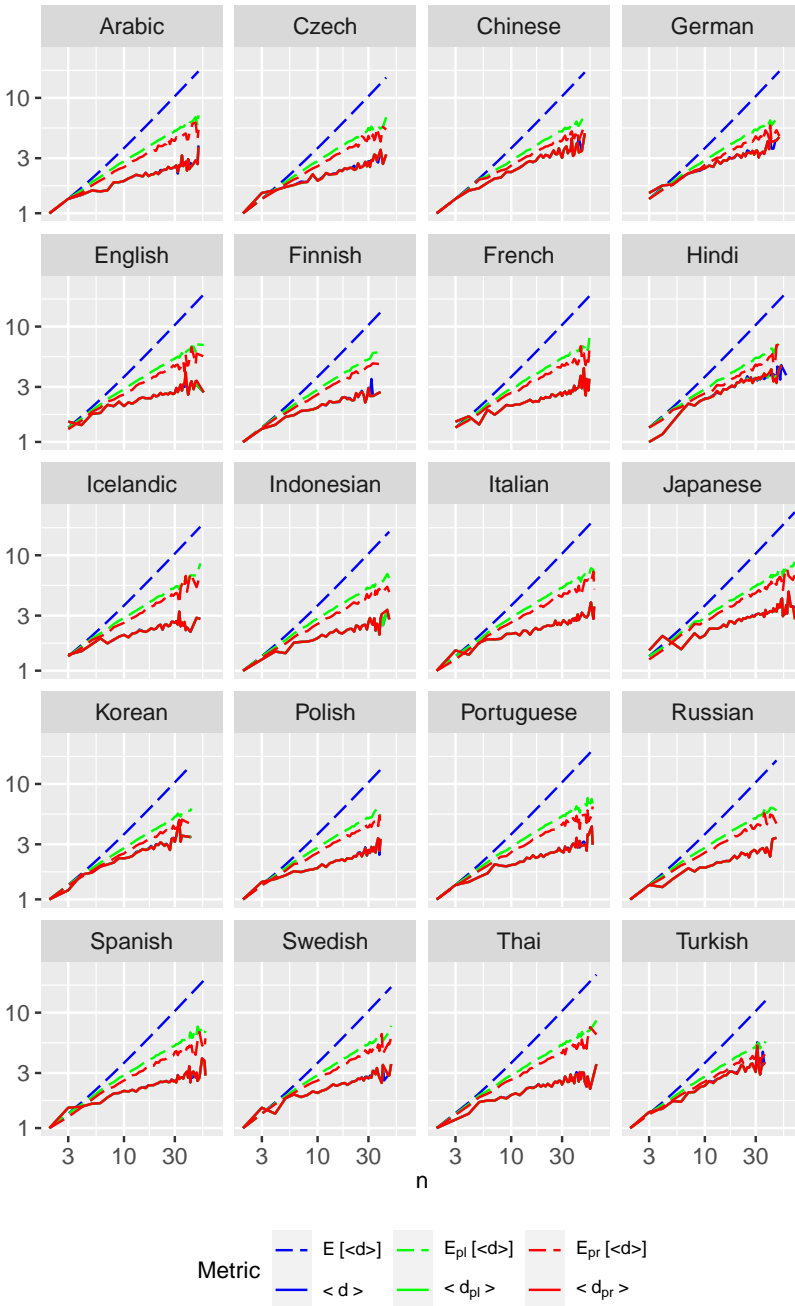


Figure 8: The scaling of  $\langle d \rangle$ , the mean dependency distance of a sentence as a function of sentence length ( $n$ ) for languages in the PSUD collection for formal constraints of increasing strength. Format is the same as in Figure 7. Again, solid lines overlap so much that only one of them can be seen in most cases

These findings are consistent between PUD and PSUD, in spite of their differences in proportions of projective and planar sentences commented above.

## 4 CONCLUSIONS AND FUTURE WORK

### 4.1 *Theory*

In Section 2.2, we have characterized planar arrangements of a given free tree  $T$  using the concept of segment (Alemany-Puig and Ferrer-i-Cancho 2022). Employing said characterization, we have shown that the number of planar arrangements of a free tree depends on its degree sequence (Proposition 1), similar to the manner in which projective arrangements of a rooted tree do (Alemany-Puig and Ferrer-i-Cancho 2022). Moreover, we have given a procedure to generate u.a.r. planar arrangements of a given free tree in Section 2.3 (Algorithm 3) which can be easily adapted to generate such arrangements exhaustively. Notably, our algorithm to generate planar arrangements is based on the generation of projective arrangements of a rooted subtree. For the sake of completeness, we have detailed a procedure to generate u.a.r. projective arrangements of a given rooted tree (Algorithm 1).

### 4.2 *Applications*

Having identified the underlying structure of planar arrangements, we have derived an arithmetic expression, in Section 2.4, for  $\mathbb{E}_{\text{pl}}[D(T)]$  (Theorem 1). We have also devised a  $O(n)$ -time algorithm to calculate this value (Proposition 1, Algorithm 4).

In Section 3, we have applied the theory developed up until that point to investigate the effect of formal constraints of increasing strength (none, planarity, projectivity) in a parallel collection and reported two main findings. First, the average dependency distance in real sentences remains practically the same while the strength of the formal constraint increases. We believe that this result stems from the high proportion of planar sentences (and the very low proportion of planar sentences that are not projective) of the PUD collection. Higher

proportions of non-planar sentences have been reported in other collections (Gómez-Rodríguez and Ferrer-i-Cancho 2017). Second, the tendency of the random baseline to have a smaller value in stronger formal constraints indicates that the strength of the dependency distance minimization effect depends on the choice of the formal constraint for the random baseline. As these formal constraints may be a side effect of dependency distance minimization (Ferrer-i-Cancho 2006; Gómez-Rodríguez and Ferrer-i-Cancho 2017; Gómez-Rodríguez *et al.* 2022; Yadav *et al.* 2022), this phenomenon suggests that

1. Formal constraints absorb the dependency distance effect.
2. A fairer evaluation of the actual degree of optimization of dependency distances or a more accurate measurement of the power of the effect of dependency distance minimization requires considering not only the magnitude of the effect with respect to some random baseline but also the formal constraint, as the latter may hide part of the dependency distance minimization effect.

In past research on syntactic dependency distance minimization,  $\mathbb{E}_{\text{pr}}[D(T^r)]$  has been the most widely used random baseline (Gildea and Temperley 2007; Liu 2008; Park and Levy 2009; Futrell *et al.* 2015). However, projectivity has a lower coverage than planarity in real sentences (Havelka 2007; Gómez-Rodríguez and g 2010). Projectivity is at risk of underestimating the strength of the dependency distance minimization principle (Ferrer-i-Cancho 2004) because of the significant reduction in the value of the random baseline (Figures 7 and 8) or the reduction of the actual dependency distances (Gómez-Rodríguez *et al.* 2022, Figure 2) that it introduces. Thanks to the research in this article, we have paved the way for replicating past research replacing  $\mathbb{E}_{\text{pr}}[D(T^r)]$  with  $\mathbb{E}_{\text{pl}}[D(T)]$ .

#### Future work

4.3

Planarity is a relaxation of projectivity but future work should address the problem of the expected value of  $D(T)$  in classes of formal constraints with even more coverage (Ferrer-i-Cancho *et al.* 2018). A promising step is the investigation of  $\mathbb{E}_{\leq k}[D(T)]$ , the expected value of  $D(T)$  conditioned to arrangements  $\pi$  such that  $C_\pi(T) \leq k$ , that is,

in arrangements such that the number of edge crossings is at most  $k$ . Notice that  $\mathbb{E}_{\leq 0}[D(T)] = \mathbb{E}_{\text{pl}}[D(T)]$ . In real languages, the average number of crossings ranges between 0.40 and 0.62 (Ferrer-i-Cancho *et al.* 2018), suggesting that  $\mathbb{E}_{\leq k}[D(T)]$  with  $k = 1$  or a small  $k$  would suffice.

## ACKNOWLEDGMENTS

We are grateful to J. L. Esteban for helpful comments and C. Gómez-Rodríguez advice on the computational linguistics literature. LAP is supported by Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the Social European Fund and a Ph.D. contract extension funded by Banco Santander. This research is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya) and the grants AGRUPS-2022 and AGRUPS-2023 from Universitat Politècnica de Catalunya.

## APPENDIX DERIVATION OF $\mathbb{E}_{\text{pr}}^{\circ}[\beta_{uv} | u]$

Here we derive the expected length of the coanchor of a (directed) edge  $uv \in E(T^u)$  in uniformly random projective arrangements of  $T^u$  conditioned to  $\pi(u) = 1$ . Following Alemany-Puig and Ferrer-i-Cancho (2022), we decompose the length of the coanchor of the (directed) edge  $uv$ ,  $\beta_{uv}$ , as the sum of the lengths of the segments in-between  $u$  and  $v$  (Figure 4). Here we use  $k_{uv}$  to denote the number of segments in-between  $u$  and  $v$ , and  $\varphi_{uv}^{(i)}$  to denote the size of the  $i$ th segment, yielding (Alemany-Puig and Ferrer-i-Cancho 2022),

$$\beta_{uv} = \sum_{i=1}^{k_{uv}} \varphi_{uv}^{(i)}.$$



By the Law of Total Expectation, we have that

$$(28) \quad \mathbb{E}_{\text{pr}}^{\circ} [\beta_{uv} | u] = \sum_{k=1}^{d(u)-1} \mathbb{E}_{\text{pr}}^{\circ} [\beta_{uv} | u, k_{uv} = k] \mathbb{P}_{\text{pr}}^{\circ} (k_{uv} = k | u),$$

where  $\mathbb{E}_{\text{pr}}^{\circ} [\beta_{uv} | u, k_{uv} = k]$  is the expectation of  $\beta_{uv}$  given that  $u$  is the root of the tree (fixed at the leftmost position), and that  $u$  and  $v$  are separated by  $k$  segments, and  $\mathbb{P}_{\text{pr}}^{\circ} (k_{uv} = k | u)$  is the probability that  $u$  and  $v$  are separated by  $k$  intermediate segments, both in uniformly random projective arrangements  $\pi$  conditioned to  $\pi(u) = 1$ , both conditioned to the root of the tree being vertex  $u$ . On the one hand,

$$(29) \quad \mathbb{E}_{\text{pr}}^{\circ} [\beta_{uv} | u, k_{uv} = k] = \mathbb{E}_{\text{pr}}^{\circ} \left[ \sum_{i=1}^k \varphi_{uv}^{(i)} | u \right] = \frac{n - s_u(v) - 1}{d(u) - 1} k.$$

Notice that this is the same result as that obtained in Alemany-Puig and Ferrer-i-Cancho 2022. Lastly, the proportion of arrangements in which the segment of  $v$  is at position  $k_{uv} + 1$  equals  $(d(u) - 1)!$ , therefore,

$$(30) \quad \mathbb{P}_{\text{pr}}^{\circ} (k_{uv} = k | u) = \frac{(d(u) - 1)! \prod_{v \in \Gamma(u)} \mathbf{N}_{\text{pr}}(T^u)}{d(u)! \prod_{v \in \Gamma(u)} \mathbf{N}_{\text{pr}}(T^u)} = \frac{1}{d(u)}.$$

Recalling that (Alemany-Puig and Ferrer-i-Cancho 2022)

$$\mathbb{E}_{\text{pr}} [\beta_{uv} | u] = \frac{s_u(u) - s_u(v) - 1}{3},$$

and plugging the results of Equations 29 and 30 into Equation 28, we get

$$\begin{aligned} \mathbb{E}_{\text{pr}}^{\circ} [\beta_{uv} | u] &= \frac{n - s_u(v) - 1}{d(u) - 1} \frac{1}{d(u)} \sum_{k=1}^{d(u)-1} k \\ &= \frac{s_u(u) - s_u(v) - 1}{2} \\ &= \frac{3}{2} \mathbb{E}_{\text{pr}} [\beta_{uv} | u]. \end{aligned}$$

## REFERENCES

- Lluís ALEMANY-PUIG, Juan Luis ESTEBAN, and Ramon FERRER-I-CANCHO (2021), The Linear Arrangement Library. A new tool for research on syntactic dependency structures, in *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, pp. 1–16, Association for Computational Linguistics, Sofia, Bulgaria, <https://aclanthology.org/2021.quasy-1.1>.
- Lluís ALEMANY-PUIG, Juan Luis ESTEBAN, and Ramon FERRER-I-CANCHO (2022), Minimum projective linearizations of trees in linear time, *Information Processing Letters*, 174:106204, doi:10.1016/j.ipl.2021.106204.
- Lluís ALEMANY-PUIG and Ramon FERRER-I-CANCHO (2022), Linear-time calculation of the expected sum of edge lengths in projective linearizations of trees, *Computational Linguistics*, 48(3):491–516, doi:10.1162/coli\_a.00442.
- Frank BERNHART and Paul C. KAINEN (1979), The book thickness of a graph, *Journal of Combinatorial Theory, Series B*, 27(3):320–331, doi:10.1016/0095-8956(79)90021-2.
- Fan R. K. CHUNG (1984), On optimal linear arrangements of trees, *Computers & Mathematics with Applications*, 10(1):43–60, doi:10.1016/0898-1221(84)90085-3.
- Thomas H. CORMEN, Charles E. LEISERSON, Ronald L. RIVEST, and Clifford STEIN (2001), *Introduction to algorithms*, The MIT Press, Cambridge, MA, USA, 2nd edition.
- Ramon FERRER-I-CANCHO (2004), Euclidean distance between syntactically linked words, *Physical Review E*, 70:056135, doi:10.1103/PhysRevE.70.056135.
- Ramon FERRER-I-CANCHO (2006), Why do syntactic links not cross?, *Europhysics Letters (EPL)*, 76(6):1228–1235, doi:10.1209/epl/i2006-10406-0.
- Ramon FERRER-I-CANCHO (2019), The sum of edge lengths in random linear arrangements, *Journal of Statistical Mechanics*, 2019(5):053401, doi:10.1088/1742-5468/ab11e2.
- Ramon FERRER-I-CANCHO and Carlos GÓMEZ-RODRÍGUEZ (2021), Anti dependency distance minimization in short sequences. a graph theoretic approach, *Journal of Quantitative Linguistics*, 28(1):50–76, doi:10.1080/09296174.2019.1645547.
- Ramon FERRER-I-CANCHO, Carlos GÓMEZ-RODRÍGUEZ, and Juan Luis ESTEBAN (2018), Are crossing dependencies really scarce?, *Physica A: Statistical Mechanics and its Applications*, 493:311–329, doi:10.1016/j.physa.2017.10.048.
- Ramon FERRER-I-CANCHO, Carlos GÓMEZ-RODRÍGUEZ, Juan Luis ESTEBAN, and Lluís ALEMANY-PUIG (2022), Optimality of syntactic dependency distances, *Physical Review E*, 105:014308, doi:10.1103/PhysRevE.105.014308.

- Ramon FERRER-I-CANCHO and Haitao LIU (2014), The risks of mixing dependency lengths from sequences of different length, *Glottology*, 5:143–155, doi:10.1515/glot-2014-0014.
- Richard FUTRELL, Kyle MAHOWALD, and Edward GIBSON (2015), Large-scale evidence of dependency length minimization in 37 languages, *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, doi:10.1073/pnas.1502134112.
- Kim GERDES, Bruno GUILLAUME, Sylvain KAHANE, and Guy PERRIER (2018), SUD or surface-syntactic universal dependencies: an annotation scheme near-isomorphic to UD, in *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 66–74, Association for Computational Linguistics, Brussels, Belgium, doi:10.18653/v1/W18-6008.
- Daniel GILDEA and David TEMPERLEY (2007), Optimizing grammars for minimum dependency length, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 184–191, Association for Computational Linguistics, Prague, Czech Republic, <https://www.aclweb.org/anthology/P07-1024>.
- David GILDEA and David TEMPERLEY (2010), Do grammars minimize dependency length?, *Cognitive Science*, 34(2):286–310, doi:10.1111/j.1551-6709.2009.01073.x.
- Carlos GÓMEZ-RODRÍGUEZ (2016), Restricted non-projectivity: Coverage vs. efficiency, *Computational Linguistics*, 42(4):809–817, doi:10.1162/COLI\_a\_00267.
- Carlos GÓMEZ-RODRÍGUEZ, Morten H. CHRISTIANSEN, and Ramon FERRER-I-CANCHO (2022), Memory limitations are hidden in grammar, *Glottometrics*, 52:39–64, doi:10.53482/2022\_52\_397.
- Carlos GÓMEZ-RODRÍGUEZ and Ramon FERRER-I-CANCHO (2017), Scarcity of crossing dependencies: a direct outcome of a specific constraint?, *Physics Review E*, 96:062304, doi:10.1103/PhysRevE.96.062304.
- Carlos GÓMEZ-RODRÍGUEZ and Joakim G (2010), A transition-based parser for 2-planar dependency structures, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1492–1501, Association for Computational Linguistics, Uppsala, Sweden, <https://aclanthology.org/P10-1151>.
- Thomas GROSS and Timothy OSBORNE (2009), Toward a practical dependency grammar theory of discontinuities, *SKY Journal of Linguistics*, 22:43–90, <http://www.linguistics.fi/julkaisut/sky2009.shtml>.
- David S. GUNDERSON (2014), *Handbook of Mathematical Induction: Theory and Applications*, Discrete Mathematics and Its Applications, CRC Press, ISBN 9781420093643, <https://www.routledgehandbooks.com/doi/10.1201/9781420093650>.

- Jiří HAVELKA (2007), Beyond projectivity: multilingual evaluation of constraints and measures on non-projective structures, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 608–615, Association for Computational Linguistics, Prague, Czech Republic, <https://aclanthology.org/P07-1077>.
- Robert A. HOCHBERG and Matthias F. STALLMANN (2003), Optimal one-page tree embeddings in linear time, *Information Processing Letters*, 87(2):59–66, doi:10.1016/S0020-0190(03)00261-8.
- Richard HUDSON (1995), Measuring syntactic difficulty, *Unpublished paper*, <https://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf>.
- Alex KRAMER (2021), Dependency lengths in speech and writing: a cross-linguistic comparison via YouDePP, a pipeline for scraping and parsing YouTube captions, in *Proceedings of the Society for Computation in Linguistics*, volume 4, pp. 359–365, doi:10.7275/pz9g-d780.
- Marco KUHLMANN and Joakim NIVRE (2006), Mildly non-projective dependency structures, in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, COLING-ACL '06, pp. 507–514, doi:10.3115/1273073.1273139.
- Haitao LIU (2008), Dependency distance as a metric of language comprehension difficulty, *Journal of Cognitive Science*, 9(2):159–191, doi:10.17791/jcs.2008.9.2.159.
- Haitao LIU, Chunshan XU, and Junying LIANG (2017), Dependency distance: a new perspective on syntactic patterns in natural languages, *Physics of Life Reviews*, 21:171–193, doi:10.1016/j.plrev.2017.03.002.
- Michael MITZENMACHER and Eli UPFAL (2017), *Probability and computing. Randomization and probabilistic techniques in algorithms and data analysis*, Cambridge University Press, ISBN 978-1-107-15488-9.
- Glyn MORRILL (2000), Incremental processing and acceptability, *Computational Linguistics*, 25(3):319–338, doi:10.1162/089120100561728.
- Joakim NIVRE (2006), Constraints on non-projective dependency parsing, in Diana MCCARTHY and Shuly WINTNER, editors, *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 73–80, Association for Computational Linguistics, Trento, Italy, <https://aclanthology.org/E06-1010>.
- Joakim NIVRE (2009), Non-projective dependency parsing in expected linear time, in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pp. 351–359, Association for Computational Linguistics, Stroudsburg, PA, USA, <https://aclanthology.org/P09-1040>.

Y. Albert PARK and Roger P. LEVY (2009), Minimal-length linearizations for mildly context-sensitive dependency trees, in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 335–343, Association for Computational Linguistics, Stroudsburg, PA, USA, <https://aclanthology.org/N09-1038/>.

Yossi SHILOACH (1979), A minimum linear arrangement algorithm for undirected trees, *SIAM Journal on Computing*, 8(1):15–32, doi:10.1137/0208002.

Daniel SLEATOR and Davy TEMPERLEY (1993), Parsing English with a link grammar, in *Proceedings of the Third International Workshop on Parsing Technologies (IWPT'93)*, pp. 277–292, ACL/SIGPARSE, <https://dblp.uni-trier.de/db/journals/corr/corr9508.html#abs-cmp-1g-9508004>.

David TEMPERLEY (2008), Dependency-length minimization in natural and artificial languages, *Journal of Quantitative Linguistics*, 15(3):256–282, doi:10.1080/09296170802159512.

David TEMPERLEY and Daniel GILDEA (2018), Minimizing syntactic dependency lengths: Typological/cognitive universal?, *Annual Review of Linguistics*, 4(1):67–80, doi:10.1146/annurev-linguistics-011817-045617.

Himanshu YADAV, Samar HUSAIN, and Richard FUTRELL (2022), Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity, *Computational Linguistics*, pp. 1–27, doi:10.1162/coli\_a\_00437.

Daniel ZEMAN, Joakim NIVRE, Mitchell ABRAMS, Elia ACKERMANN, Noëmi AEPLI, Željko AGIĆ, Lars AHRENBERG, Chika Kennedy AJEDE, Gabrielé ALEKSANDRAVIČIŪTĖ, Lene ANTONSEN, Katya APLONOVA, Angelina AQUINO, Maria Jesús ARANZABE, Gashaw ARUTIE, Masayuki ASAHARA, Luma ATEYAH, Furkan ATMACA, Mohammed ATTIA, Aitziber ATUTXA, Liesbeth AUGUSTINUS, Elena BADMAEVA, Miguel BALLESTEROS, Esha BANERJEE, Sebastian BANK, Verginica BARBU MITITELU, Victoria BASMOV, Colin BATCHELOR, John BAUER, Kepa BENGOETXEA, Yevgeni BERZAK, Irshad Ahmad BHAT, Riyaz Ahmad BHAT, Erica BIAGETTI, Eckhard BICK, Agnė BIELINSKIENĖ, Rogier BLOKLAND, Victoria BOBICEV, Loïc BOIZOU, Emanuel BORGES VÖLKER, Carl BÖRSTELL, Cristina BOSCO, Gosse BOUMA, Sam BOWMAN, Adriane BOYD, Kristina BROKAITĖ, Aljoscha BURCHARDT, Marie CANDITO, Bernard CARON, Gauthier CARON, Tatiana CAVALCANTI, Gülşen CEBIROĞLU ERYIĞIT, Flavio Massimiliano CECCHINI, Giuseppe G. A. CELANO, Slavomír ČĚPLŮ, Savas CETIN, Fabricio CHALUB, Ethan CHI, Jinho CHOI, Yongseok CHO, Jayeol CHUN, Alessandra T. CIGNARELLA, Silvie CINKOVÁ, Aurélie COLLOMB, Çağrı ÇÖLTEKIN, Miriam CONNOR, Marine COURTIN, Elizabeth DAVIDSON, Marie-Catherine DE MARNEFFE, Valeria DE PAIVA, Elvis DE SOUZA, Arantza DIAZ DE ILARRAZA, Carly DICKERSON, Bamba DIONE, Peter DIRIX, Kaja DOBROVOLJC, Timothy DOZAT, Kira DROGANOVA, Puneet DWIVEDI, Hanne

ECKHOFF, Marhaba ELI, Ali ELKAHKY, Binyam EPHREM, Olga ERINA, Tomaž ERJAVEC, Aline ETIENNE, Wograine EVELYN, Richárd FARKAS, Hector FERNANDEZ ALCALDE, Jennifer FOSTER, Cláudia FREITAS, Kazunori FUJITA, Katarína GAJDOŠOVÁ, Daniel GALBRAITH, Marcos GARCIA, Moa GÄRDENFORS, Sebastian GARZA, Kim GERDES, Filip GINTER, Iakes GOENAGA, Koldo GOJENOLA, Memduh GÖKIRMAK, Yoav GOLDBERG, Xavier GÓMEZ GUINOVART, Berta GONZÁLEZ SAAVEDRA, Bernadeta GRICIŪTĖ, Matias GRIONI, Loïc GROBOL, Normunds GRŪZĪTIS, Bruno GUILLAUME, Céline GUILLOT-BARBANCE, Tunga GÜNGÖR, Nizar HABASH, Jan HAJIČ, Jan HAJIČ JR., Mika HÄMÄLÄINEN, Linh HÀ MỸ, Na-Rae HAN, Kim HARRIS, Dag HAUG, Johannes HEINECKE, Oliver HELLWIG, Felix HENNIG, Barbora HLADKÁ, Jaroslava HLAVÁČOVÁ, Florinel HOCIUNG, Petter HOHLE, Jena HWANG, Takumi IKEDA, Radu ION, Elena IRIMIA, Ọlájídé ISHOLA, Tomáš JELÍNEK, Anders JOHANNSEN, Hildur JÓNSDÓTTIR, Fredrik JØRGENSEN, Markus JUUTINEN, Hüner KAŞIKARA, Andre KAASEN, Nadezhda KABAEVA, Sylvain KAHANE, Hiroshi KANAYAMA, Jenna KANERVA, Boris KATZ, Tolga KAYADELEN, Jessica KENNEY, Václava KETTNEROVÁ, Jesse KIRCHNER, Elena KLEMENTIEVA, Arne KÖHN, Abdullatif KÖKSAL, Kamil KOPACEWICZ, Timo KORAKIANGAS, Natalia KOTSYBA, Jolanta KOVALEVSKAITĖ, Simon KREK, Sookyoung KWAK, Veronika LAIPPALA, Lorenzo LAMBERTINO, Lucia LAM, Tatiana LANDO, Septina Dian LARASATI, Alexei LAVRENTIEV, John LEE, Phưởng LÊ HỒNG, Alessandro LENCI, Saran LERTPRADIT, Herman LEUNG, Maria LEVINA, Cheuk Ying LI, Josie LI, Keying LI, KyungTae LIM, Yuan LI, Nikola LJUBEŠIĆ, Olga LOGINOVA, Olga LYASHEVSKAYA, Teresa LYNN, Vivien MACKETANZ, Aibek MAKAZHANOV, Michael MANDL, Christopher MANNING, Ruli MANURUNG, Cătălina MĂRĂNDUC, David MAREČEK, Katrin MARHEINECKE, Héctor MARTÍNEZ ALONSO, André MARTINS, Jan MAŠEK, Hiroshi MATSUDA, Yuji MATSUMOTO, Ryan McDONALD, Sarah MCGUINNESS, Gustavo MENDONÇA, Niko MIEKKA, Margarita MISIRPASHAYEVA, Anna MISSILÄ, Cătălin MITITELU, Maria MITROFAN, Yusuke MIYAO, Simonetta MONTEMAGNI, Amir MORE, Laura MORENO ROMERO, Keiko Sophie MORI, Tomohiko MORIOKA, Shinsuke MORI, Shigeki MORO, Bjartur MORTENSEN, Bohdan MOSKALEVSKYI, Kadri MUISCHNEK, Robert MUNRO, Yugo MURAWAKI, Kaili MÜÜRİSEP, Pinkey NAINWANI, Juan Ignacio NAVARRO HORŃIACEK, Anna NEDOLUZHKO, Gunta NEŠPORE-BĚRZKALNE, Lương NGUYỄN THỊ, Huỳen NGUYỄN THỊ MINH, Yoshihiro NIKAIDO, Vitaly NIKOLAEV, Rattima NITISAROJ, Hanna NURMI, Stina OJALA, Atul Kr. OJHA, Adédayò OLÚÒKUN, Mai OMURA, Emeka ONWUEGBUZIA, Petya OSENOVA, Robert ÖSTLING, Lilja ØVRELID, Şaziye Betül ÖZATEŞ, Arzucan ÖZGÜR, Balkız ÖZTÜRK BAŞARAN, Niko PARTANEN, Elena PASCUAL, Marco PASSAROTTI, Agnieszka PATEJUK, Guilherme PAULINO-PASSOS, Angelika PELJAK-ŁAPIŃSKA, Siyao PENG, Cenel-Augusto PEREZ, Guy PERRIER, Daria PETROVA, Slav PETROV, Jason PHELAN, Jussi PIITULAINEN, Tommi A

PIRINEN, Emily PITLER, Barbara PLANK, Thierry POIBEAU, Larisa PONOMAREVA, Martin POPEL, Lauma PRETKALNIŅA, Sophie PRÉVOST, Prokopis PROKOPIDIS, Adam PRZEPIÓRKOWSKI, Tiina PUOLAKAINEN, Sampo PYYSALO, Peng QI, Andriela RÄÄBIS, Alexandre RADEMAKER, Loganathan RAMASAMY, Taraka RAMA, Carlos RAMISCH, Vinit RAVISHANKAR, Livy REAL, Petru REBEJA, Siva REDDY, Georg REHM, Ivan RIABOV, Michael RIESSLER, Erika RIMKUTĚ, Larissa RINALDI, Laura RITUMA, Luisa ROCHA, Mykhailo ROMANENKO, Rudolf ROSA, Valentin ROȘCA, Davide ROVATI, Olga RUDINA, Jack RUETER, Shoal SADDE, Benoît SAGOT, Shadi SALEH, Alessio SALOMONI, Tanja SAMARDŽIĆ, Stephanie SAMSON, Manuela SANGUINETTI, Dage SÁRG, Baiba SAULĪTE, Yanin SAWANAKUNANON, Salvatore SCARLATA, Nathan SCHNEIDER, Sebastian SCHUSTER, Djamé SEDDAH, Wolfgang SEEKER, Mojgan SERAJI, Mo SHEN, Atsuko SHIMADA, Hiroyuki SHIRASU, Muh SHOHIBUSSIRRI, Dmitry SICHINAVA, Aline SILVEIRA, Natalia SILVEIRA, Maria SIMI, Radu SIMIONESCU, Katalin SIMKÓ, Mária ŠIMKOVÁ, Kiril SIMOV, Maria SKACHEDUBOVA, Aaron SMITH, Isabela SOARES-BASTOS, Carolyn SPADINE, Antonio STELLA, Milan STRAKA, Jana STRNADOVÁ, Alane SUHR, Umot SULUBACAK, Shingo SUZUKI, Zsolt SZÁNTÓ, Dima TAJI, Yuta TAKAHASHI, Fabio TAMBURINI, Takaaki TANAKA, Samson TELLA, Isabelle TELLIER, Guillaume THOMAS, Liisi TORGA, Marsida TOSKA, Trond TROSTERUD, Anna TRUKHINA, Reut TSARFATY, Utku TÜRK, Francis TYERS, Sumire UEMATSU, Roman UNTILOV, Zdeňka UREŠOVÁ, Larraitz URIA, Hans USZKOREIT, Andrius UTKA, Sowmya VAJJALA, Daniel VAN NIEKERK, Gertjan VAN NOORD, Viktor VARGA, Eric VILLEMONTÉ DE LA CLERGERIE, Veronika VINCZE, Aya WAKASA, Lars WALLIN, Abigail WALSH, Jing Xian WANG, Jonathan North WASHINGTON, Maximilan WENDT, Paul WIDMER, Seyi WILLIAMS, Mats WIRÉN, Christian WITTERN, Tsegay WOLDEMARIAM, Tak-sum WONG, Alina WRÓBLEWSKA, Mary YAKO, Kayo YAMASHITA, Naoki YAMAZAKI, Chunxiao YAN, Koichi YASUOKA, Marat M. YAVRUMYAN, Zhuoran YU, Zdeněk ŽABOKRTSKÝ, Amir ZELDES, Hanzhi ZHU, and Anna ZHURAVLEVA (2020), Universal Dependencies 2.6, <http://hdl.handle.net/11234/1-3226>, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

*Lluís Alemany-Puig*

ORCID 0000-0002-3874-991X

lluis.alemany.puig@upc.edu

Quantitative, Mathematical and  
Computational Linguistics Research  
Group. Departament de Ciències de la  
Computació, Universitat Politècnica  
de Catalunya (UPC), Barcelona,  
Catalonia, Spain

*Ramon Ferrer-i-Cancho*

ORCID 0000-0002-7820-923X

rferrericancho@cs.upc.edu

Quantitative, Mathematical and  
Computational Linguistics Research  
Group. Departament de Ciències de la  
Computació, Universitat Politècnica  
de Catalunya (UPC), Barcelona,  
Catalonia, Spain

Lluís Alemany-Puig and Ramon Ferrer-i-Cancho (2024), *The expected sum of edge lengths in planar linearizations of trees*, *Journal of Language Modelling*, 12(1):1–42

DOI <https://dx.doi.org/10.15398/jlm.v12i1.362>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

CC BY <http://creativecommons.org/licenses/by/4.0/>



# Detecting inflectional patterns for Croatian verb stems using class activation mapping

*Domagoj Ševerdija<sup>1</sup>, Rebeka Čorić<sup>1</sup>, Marko Orešković<sup>2</sup>,  
and Lucian Šošić<sup>3</sup>*

<sup>1</sup> Department of Mathematics, University J. J. Strossmayer of Osijek

<sup>2</sup> National and University Library in Zagreb

<sup>3</sup> Faculty of Humanities and Social Sciences in Split

## ABSTRACT

All verbal forms in the Croatian language can be derived from two basic forms: the infinitive and the present stems. In this paper, we present a neural computation model that takes a verb in an infinitive form and finds a mapping to a present form. The same model can be applied vice-versa, i.e. map a verb from its present form to its infinitive form. Knowing the present form of a given verb, one can deduce its inflections using grammatical rules. We experiment with our model on the Croatian language, which belongs to the Slavic group of languages. The model learns a classifier through these two classification tasks and uses class activation mapping to find characters in verbs contributing to classification. The model detects patterns that follow established grammatical rules for deriving the present stem form from the infinitive stem form and vice-versa. If mappings can be found between such slots, the rest of the slots can be deduced using a rule-based system.

*Keywords:*  
*Croatian infinitive  
and present verb  
stems,  
convolutional  
neural network,  
class activation  
mapping*

Inflection of verbs in morphologically rich languages such as Croatian is important as it conveys grammatical information such as tense, aspect, mood, and voice, making sentences shorter and more precise. Computing the proper inflections of a verb is a well-established problem in computational morphology. Finding pairs of basic forms of Croatian verbs can convey enough information to apply the appropriate inflection rules. There is no way of setting the pairs right without applying brute force, i.e., checking each pair for each particular verb. Given the number of verbs in Croatian, checking manually can be time-consuming. Finding a mapping between basic verb forms can automate this process.

The choice of the proper verb form is in many ways arbitrary, as are most grammatical features. Many seemingly similar verbs belong to different conjugational paradigms for no apparent reason. Were it not so, creating a verb generator would be trivial. Instead, a sample of verbs can be used as input in order to obtain some statistical indicators on the likelihood of verbs with certain phonetical features appearing in certain conjugation classes, e.g.:

- Verbs ending in *-ati* that are derived from nouns or adjectives (*pilati*, *brzati*) have a strong tendency to take the present tense ending *-am*.
- Verbs ending in *-ati* that are loanwords (*krcati*, *peglati*) have a strong tendency to take the present tense ending *-am*. Searching for atypical phoneme clusters will thus be a statistical indicator of the ending.
- Verbs in *-ovati* take the present tense ending *-ujem*, except for some very short ones (*lovati*→*lovam*).

This view is by no means oversimplified; we do not suggest that the language is regular. There are irregularities such as *vreti*, which has an anomalous 3rd person plural. These anomalies in personal endings are very rare, however, and can be enumerated manually. The main goal of this paper is therefore to derive a model to find proper pairs of basic forms of Croatian verbs, conveying enough information to apply the appropriate inflection rules.

First, we shall describe our problem from a (broader) linguistic perspective and then give an overview of the current state-of-the-art computational approaches for tackling this problem. The main contribution is given in Section 2, with an example of rule-based system application in Section 3.4.

### The Croatian verb system

1.1

The Croatian verb system is inherited from Proto-Slavic and conservatively preserves the ancient inflection. Consequently, the phonological and morphological rules governing conjugation are often opaque. There are six verb categories: **person**, **number**, **tense**, **mood**, **voice** and **aspect**. The first five categories are common in all Indo-European languages that preserve verbal inflection. Aspect can be either **perfective** or **imperfective**,<sup>1</sup> thus individual verbs are almost always either perfective or imperfective (a few can be biaspectual). Verb forms can be both finite and non-finite. Finite forms are conjugated in person and number. There are four moods: **indicative**, **imperative**, **optative**, and **conditional**, as described in Barić *et al.* 2005. The indicative mood has seven finite tenses: **present**, **perfect**, **orist**, **imperfect**, **pluperfect**, **future**, and **perfective future**. Conditional has a present and a perfect form. This gives a total of 11 finite forms (Table 1).

	Present	Perfect	Aorist	Imperfect	Pluperfect	Future	Perfective Future
Indicative	•	•	•	•	•	•	•
Imperative	•						
Optative	•						
Conditional	•	•					

Table 1:  
Finite verb forms

<sup>1</sup> Croatian aspects are referred to as “perfective” and “imperfective” in English. “Perfect” and “imperfect” are tenses. The traditional names are derived from similar roots, but are not the same. In fact, perfect tense can be both perfective and imperfective. Imperfect, however, is always imperfective.

Table 2:  
Thematic  
suffixes with  
final morpheme

Verb form	Stem	Thematic Suffix	Final morpheme
<i>raditi</i> 'to work'	<i>rad-</i>	<i>-i-</i>	<i>-ti</i>
<i>radim</i> 'I work'	<i>rad-</i>	<i>-i-</i>	<i>-m</i>
<i>radih</i> 'I worked'	<i>rad-</i>	<i>-i-</i>	<i>-h</i>
<i>tresti</i> 'to shake'	<i>tres-</i>	<i>-ø-</i>	<i>-ti</i>
<i>tresem</i> 'I shake'	<i>tres-</i>	<i>-e-</i>	<i>-m</i>
<i>potresoh</i> 'I shook'	<i>potres-</i>	<i>-o-</i>	<i>-h</i>

Non-finite verb forms (i.e. the **infinitive** and the **participles**) are not conjugated in person or number. Participles are not conjugated, but they can be declined as adjectives. They do not have standardized English names: here, they will be referred to as the *l*-participle, passive participle, present participle, and past participle. This gives a total of 5 non-finite forms. All of these forms can be synthetic or analytic. Synthetic forms are primary forms that consist of a single word. Analytic forms are derived by combining synthetic forms and auxiliary verbs. For example, the perfect tense is constructed using the *l*-participle and the auxiliary verb *biti* 'to be' in the present tense. The following forms are analytic: perfect, pluperfect, future, perfective future, optative, present conditional, and past conditional. The remaining 9 forms are synthetic: present, aorist, imperfect, imperative, *l*-participle, passive participle, present participle, past participle, and infinitive. Given that analytic forms can be derived from synthetic forms, describing the Croatian verb system can be reduced to deriving these 9 forms.

All synthetic forms are constructed from their base forms (stems). Stems are further modified with affixes (prefixes or suffixes) to produce different verb forms. Only suffixes are used to produce verbal conjugational forms. To describe the Croatian verb system, it is vital to properly parse the verbs – and identify which suffixes and stems exist. For an example of parsing, see the Table 2.

The purpose of Table 2 is to illustrate that stems can receive multiple morphemes, of two different types. Final suffixes, depend on the exact verb form (e.g. 1st person singular present). Each form has its characteristic suffix, uniform across the conjugations.<sup>2</sup> The thematic

<sup>2</sup>There are several irregularities in common verbs, e.g. 1st person singular *moгу* and *hoću*, and *l*-participle *išao*.

suffix, on the other hand, varies between different classes of verbs, as will be demonstrated in Section 1.2. The next task is to identify the stems. Nine synthetic verb forms referred to in Section 1.1 can all be considered to have their own stem, as discussed in Silić and Pranjković 2005. Such a situation is rather complex; fortunately, some of these stems are derived from others. For example, the present participle stem can be trivially derived from the present stem: it is always identical to 3rd person plural present tense followed by the suffix *-ći*. While other stems require more complex derivational rules, a regular derivation is still possible. In fact, all the stems can be regularly derived from two basic stems: the present and the infinitive. This is the “principal problem” of the “Slavic verbal system” (Micklesen 1974). The problem is referred to as “Slavic” since modern Slavic languages have all preserved old conjugational classes up to a point.

Historically, three types of verb classifications have been developed for the Slavic languages: infinitive stem, present stem, and basic stem classifications (Mihaljević 2014).

Infinitive stem classifications are the oldest, dating back to Dobrovský (1809) who divided verbs according to the thematic suffix preceding the infinitive suffix *-ti*. To further describe the verbs, however, the present stem was required, so Dobrovský divided his first class into three groups (A, B, and C). This system was soon adapted for other Slavic languages. Present stem classifications prioritize present stems over infinitive stems when classifying verbs. However, the distinction is merely hierarchical, as both forms are still required to conjugate the verb properly.

Basic stem classifications were devised with an intent to derive the entire conjugation from a single stem. However, they still require knowing whether the basic stem is the present or the infinitive to work properly.<sup>3</sup>

Thus, given the knowledge of the present and the infinitive stems, the remaining Croatian verb forms can be derived regularly (minus the few irregularities in common verb forms, as stated above).

---

<sup>3</sup>For example, the classification of OCS (Old Church Slavic) verbs by Lunt (2001) differentiates between nine classes, each defined by a single classifier. The classifier is either the infinitive thematic suffix (for six classes) or the present suffix (for the remainder).

## 1.2 *Deriving the present from the infinitive and vice versa*

The first item which must be taken into account when deriving the present from the infinitive and vice versa is the **present thematic suffix**. As explained above, the thematic suffix is the suffix preceding the present final suffixes (*-m, -š, -ø, -mo, -te, -e/u*). The personal suffixes are always the same, so it is the thematic suffix that permits the existence of multiple present classes.

With regard to the **present ending (thematic suffix + personal suffix)**, up to five main groups are identified by grammarians: athematic, *-em, -im, -am*, and *-jem* present groups.

The *-em* and *-im* groups are fully regular. The athematic group can be considered irregular and, in standard Croatian, it consists of only the verb *biti* 'be'. The *-am* group is a contracted form of the *-em* group (e.g. *pěvajem*→*pjevam*/*I sing*), and the *-jem* group can be considered a subset of the *-em* group as well. However, *-am, -em, and -jem* verbs will be analyzed separately in this paper, in accordance with Croatian linguistic practice.

The other vital feature is the **infinitive thematic suffix**. It precedes the infinitive suffix *-ti*, and most grammarians isolate six infinitive thematic suffixes (*-ø, -nø, -ě, -i, -a* and *-ova*). The *-ø* suffix causes complex shifts, so verbs with various endings (*-eti, -sti, -rti, -ći*) shall be considered here. Besides the present and infinitive suffixes, there are further sound shifts (like ablaut) that render the conjugation less predictable. These will not be addressed herein, but the interested reader can consult Silić and Pranjković 2005.

## 1.3 *Previous work*

Over the past decade, the popularity of supervised methods has produced computational inflectional models for several morphologically-rich languages (see Durrett and DeNero 2013, Barros *et al.* 2017, and Dinu *et al.* 2012 and references therein).

There is a body of work that tries to give morphological transducers in the form of software components that performs morphological generation (e.g. for the Tulu language in Antony *et al.* 2012, the Hindi language in Goyal and Lehal 2008, the German language in Zielinski

*et al.* 2009, Arabic languages in Habash and Rambow 2006 or for the Russian and the Ukrainian language in Korobov 2015). Most of the work uses some form of rule-based patterns which are defined by experts and are specific for each language. The algorithm then uses root words and appends suffixes based on these rules. In such an approach, each part of speech (e.g. nouns, verbs, or adjectives) has a different set of rules, which makes this approach very exhaustive.

For the Croatian language, there is a morphological generator actively being developed and used within the Croatian Online Syntactic and Semantic Framework described in Orešković *et al.* 2016.

In his PhD thesis, Wicentowski (2002) developed a minimally supervised framework of methods (combining supervised and unsupervised methods) for multilingual inflectional morphology covering 32 languages, but not including Croatian language, for the purposes of lemmatization. It is also worth noting that SIGMORPHON<sup>4</sup> (the Special Interest Group on Computational Morphology and Phonology) is a series of workshops and shared tasks focused on computational analysis of word structure in different languages, aiming to develop models that can generate word forms given linguistic information. It promotes research in computational morphology and phonology and in a recent shared task, SIGMORPHON 2023,<sup>5</sup> they asked the participants to design a model that learns to generate morphological inflections from a lemma and a set of morphosyntactic features of the target form for a broad range of languages. Each language in the task had its own training, development, and test datasets, but the Croatian was not provided. They also provided baselines (non-neural and neural models) for comparison.

### *Our contribution*

1.4

The main contribution of the paper is a convolutional neural network model that takes Croatian verbs in infinitive stem form and classifies them into the appropriate present stem form and vice-versa. The classifier provides information that can be used by a transducer to

---

<sup>4</sup><https://sigmorphon.github.io/>

<sup>5</sup><https://github.com/sigmorphon/2023InflectionST>

compute the proper inflection of a given verb. Moreover, the model highlights feature maps that “voted” for proper classification, i.e. it highlights all the characters within a verb that were significant for classification. This is in line with the contemporary attempt to have “explainable” AI models (xAI) (for more information on xAI in NLP see Danilevsky *et al.* 2020). Compared to Wicentowski 2002, our model is relatively simple without any explicit feature design and supervision. On the other hand, the shared task “Part 1: Typologically Diverse Morphological (Re-)Inflection” from SIGMORPHON 2023 does provide a general framework of deriving inflections from a given lemma as an end-to-end system. In our setup, the model is used as an aid to the rule-based parser.

## 2 STEM MAPPING COMPUTATION

### 2.1 *The model*

We propose a neural-network-based computation model that learns to map Croatian verbs from the infinitive stem form to the present stem form and vice-versa, as described in Section 1.2. We refer to the former as INF2PRES and to the latter as PRES2INF problem. It is considered a classification problem. Our model is essentially a convolutional neural network and Section 3.3 empirically examines the appropriateness of such architecture.

As input, our model takes a single verb  $x = \langle c_1, c_2, \dots, c_n \rangle$  as a sequence of  $n$  characters  $c_i$  in infinitive and 1st person singular form respectively. Characters  $c_i$  are taken from a predefined alphabet  $V$  of bounded size and assigned a unique symbolic representation  $c_i \in \{0, 1\}^{|V|}$  (1-hot encoding). In our model, we use embedding  $\mathbf{c}_i = \mathbf{E}c_i$  to compute dense vector representation of  $c_i$ , where  $\mathbf{E} \in \mathbb{R}^{d_e \times |V|}$  and  $d_e$  is an embedding dimension. For a window of characters

$$\mathbf{x}_i = \mathbf{c}_{i:i+k_r-1} := \langle \mathbf{c}_i, \mathbf{c}_{i+1}, \dots, \mathbf{c}_{i+k_r-1} \rangle$$

of size  $k_r$ , we apply a total of  $K$   $l$ -channel 1D convolutions of size  $k_r$ ,  $r = 1, 2, \dots, K$ , which produces a feature vector:

$$\mathbf{f}_i^{(r)} = g(\mathbf{U}^{(r)}\mathbf{x}_i + \mathbf{b}^{(r)}) \in \mathbb{R}^l$$



for  $i = 1, 2, \dots, m_r$ , where  $\mathbf{U}^{(r)} \in \mathbb{R}^{l \times k_r \times d_e}$ ,  $\mathbf{b}^{(r)} \in \mathbb{R}^l$ . Therefore,  $\mathbf{f}_{1:m_r}^{(r)}$  defines a feature map for an input  $x$  with respect to the  $r$ -th convolution. A wide convolution is applied: before the application of the filters, zero padding, if needed, is added before the first and after the last element of  $\mathbf{x}_i$ , making sure that the number of times that each character is included in the receptive field during convolution is the same, irrespective of the character's position in the word. Therefore,  $m_r = n + k_r - 1$ . A ReLU activation function is used for every convolutional layer:  $g = \text{ReLU}$ .

Every vector  $\mathbf{f}_{1:m_r}^{(r)}$  obtained from a convolutional layer is max-pooled, which results in  $K \times l$  1-dimensional vectors. These vectors are concatenated into a new vector as  $\mathbf{z} \in \mathbb{R}^{Kl}$ , which is then relayed to a fully connected layer that outputs a score vector of dimension equal to the number of classes  $c$ :

$$(1) \quad \mathbf{y}(x) = \mathbf{W}_{f_c} \mathbf{z} + \mathbf{b}_{f_c},$$

where  $\mathbf{W}_{f_c} \in \mathbb{R}^{c \times Kl}$ ,  $\mathbf{b}_{f_c} \in \mathbb{R}^c$ . A softmax normalization is applied to vector  $\mathbf{y}$  giving a probability vector over all classes

$$[P(x \in C_i)]_{i=1,2,\dots,c} = \text{softmax}(\mathbf{y}(x)) \in [0, 1]^c.$$

Before the linear layer, a dropout technique is used as a regularization method. The index of a maximum value in the resulting vector is the ordinal number of a class, namely,  $C(x)$  to which verb  $x$  should be classified, thus:

$$C(x) = \arg \max_{i \in \{1,2,\dots,c\}} P(x \in C_i).$$

See Figure 1 for an illustration. For the sake of simplicity, we denote our model as a function  $C(x) = \text{CNN}_{\Theta}(x)$ , where  $\Theta$  are learned parameters for the model during training.

### Class activation mapping

2.2

To obtain information about which characters contributed the most to the classification, class activation mapping (CAM) was used, as described in Lee *et al.* (2018). The main idea is as follows: for a given verb  $x$ , we compute a predicted class  $C(x) = \text{CNN}_{\Theta}(x)$  and look for

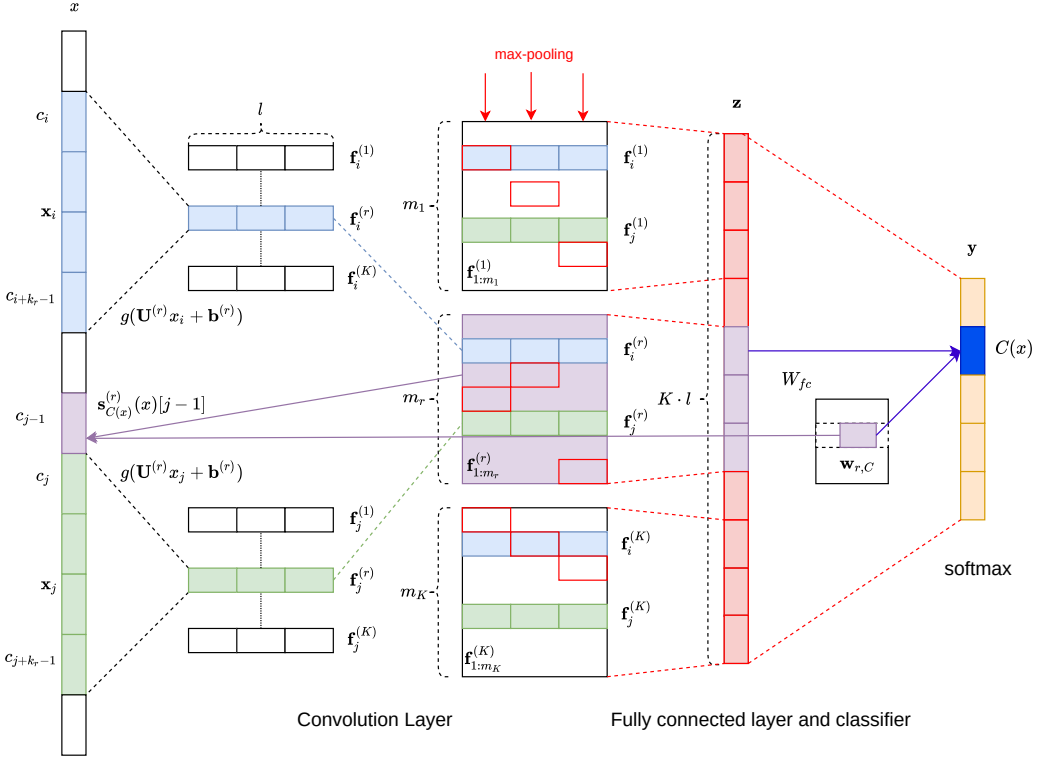


Figure 1: Two different windows of  $x$ , showing how they are used in convolution layers. The final output  $C(x)$  is a predicted class for  $x$  and it is computed using all the available windows of  $x$ . The backward arrows indicate how the CAM filter was computed to score a contribution from  $c_{j-1}$  for the classification of  $x$  to class  $C(x)$  with respect to the  $r$ -th convolution

characters of  $x$  whose feature maps are significant for the classification.

Let us rewrite (1) to consider the contribution of  $r$ -th convolution with  $l$  channels to the  $C = i$  class (without loss of generality, we omit  $\mathbf{b}_{fc}$  and assume  $C \in \{1, 2, \dots, c\}$ ):

$$y(x)[i] = (\mathbf{W}_{fc}\mathbf{z})[i] = \sum_{r=1}^K \sum_{j=(r-1)l+1}^{rl} \mathbf{W}_{fc}[i, j]\mathbf{z}[j].$$

Note that  $\mathbf{z}$  is a pooled vector by construction, we therefore want to apply the same weights on the entire feature map of  $r$ -th convo-

lution, with  $\mathbf{w}_{r,C} = W_{fc}[i, (r-1)l + 1 : rl]$ . Thus, we end up with a vector:

$$\mathbf{v}_C^{(r)} = \mathbf{f}_{1:m_r}^{(r)} \mathbf{w}_{r,C}^\top.$$

Unfortunately,  $\mathbf{v}_C^{(r)}$  is a  $m_r$ -dimensional vector and depends on the type of convolution used, but it can be reduced to a fixed-size vector whose size is independent of the convolution. We achieve this using max-pooling with a window of size  $k_r$  and step 1 over  $\mathbf{v}_C^{(r)}$  deriving a vector:

$$\mathbf{s}_C^{(r)} = \left[ \max \left( \mathbf{v}_C^{(r)} [p : p + k_r - 1] \right) \right]_{p=1,2,\dots,n} \in \mathbb{R}^n.$$

The CAM returns a score for every character in  $x$  contributing to class  $C$  over all convolutions as:

$$\text{CAM}(x, C) = \sum_{r=1}^K \mathbf{s}_C^{(r)}.$$

An illustration of CAM computation is shown in Figure 1. Examples of CAM application can be seen in Figure 4 (see page 57) and Figure 5 (see page 58).

## EXPERIMENTS

3

### Dataset

3.1

Our model was trained and evaluated on a set of Croatian verbs extracted from several lexical resources such as the Croatian WordNet (CroWN) described in Raffaelli *et al.* 2008, the Croatian linguistic portal (HJP),<sup>6</sup> and CroDerIV,<sup>7</sup> the Croatian lexicon of lexical and derivational morphemes by Šojat *et al.* (2012). The first resource was parsed using the NLTK<sup>8</sup> interface for the Open Multilingual Wordnet<sup>9</sup> by searching English verb synsets and retrieving lemmas in the Croatian language. These lemmas were used to query the CroDerIV and HJP

<sup>6</sup><http://hjp.znanje.hr>

<sup>7</sup><http://croderiv.ffzg.hr/Croderiv>

<sup>8</sup><https://www.nltk.org>

<sup>9</sup><https://omwn.org/omw1.html>

Table 3:  
INF2PRES  
statistical  
overview

Classes	Train	Val	Test	Whole corpus
<i>-am</i>	36.04	35.96	35.95	36.02
<i>-im</i>	34.79	34.81	34.76	34.79
<i>-jem</i>	11.32	11.33	11.39	11.33
<i>-em</i>	17.85	17.90	17.90	17.86
Corpus breakdown	80.95	9.03	10.02	100.00

Table 4:  
PRES2INF  
statistical  
overview

Classes	Train	VTest	Whole corpus	
<i>-ati</i>	53.82	53.42	53.23	53.73
<i>-iti</i>	32.12	31.76	31.82	32.05
<i>-jeti</i>	3.34	3.42	3.37	3.35
<i>-eti</i>	0.60	0.65	0.73	0.62
<i>-uti</i>	6.09	6.03	6.16	6.09
<i>-sti</i>	2.09	2.12	2.20	2.10
<i>-rti</i>	0.31	0.65	0.59	0.37
<i>-ĉi</i>	1.63	1.95	1.91	1.69
Corpus breakdown	80.80	9.10	10.10	100.00

search engines to obtain present stem forms. In addition, we queried HJP for verbs not found in CroWN and added them to the dataset.

The total number of verbs collected from these resources was 6794, manually organized as infinitive and present stem pairs for each verb. All pairs were verified by a human annotator.

The dataset for training and evaluation was organized as pairs  $(x_{inf}, x_{pres})$ , where  $x_{inf}$  denotes a verb in infinitive form and  $x_{pres}$  a 1st person singular present form. These forms are represented with appropriate suffixes as described in Section 1.2. The total number of available verbs is partitioned into train, validation, and test datasets with an 80:10:10 split by random sampling without replacement. A statistical overview of our dataset expressed as percentages is given in tabular form. Classes represent verb suffixes for the 1st person singular (Table 3) and infinitive form (Table 4). The most numerous classes are *-am* and *-im*, covering over 60% of the verbs for the INF2PRES and *-ati* and *-iti* covering over 80% verbs in the PRES2INF classification problem.

The model was implemented in PyTorch ver. 1.8.0 (Paszke *et al.* 2019) and deployed on AMD Zen 12 CPU with 64GB of RAM and GeForce 2070 RTX GPU. The training time for both classification tasks took less than one minute per epoch and was trained for 30 epochs using the ADAM optimizer described in Kingma and Ba 2015. For initial character embeddings, we used FastText from Bojanowski *et al.* 2017, which gave slightly better results than random initialization.

Several hyperparameters of the  $\text{CNN}_\Theta$  model had to be tuned before testing the model. These parameters were the number of filters, filter sizes, and dropout rate. Batch size and learning rate were also tuned for the training process. Hyperparameter tuning was conducted by exploring different values of parameters with 10-fold cross-validation. Parameters yielding the best average loss on validation sets were used to train the model. Hyperparameter tuning showed that, for both types of classifications, the same parameters can be used. The resulting parameters for the model can be seen in Table 5. The code is publicly available at a GitHub repository.<sup>10</sup>

Parameter	Value
l (number of filters)	36
filter sizes ( $k_r$ )	1, 2, 3, 5
dropout rate	0.1
batch size	50
learning rate	0.005

Table 5:  
Hyperparameters used  
for training the classifier

Table 6 reports the classification performance for our model on the test dataset for our model, in terms of accuracy and micro/macro/weighted  $F_1$  scores. In both classification tasks, the model achieved relatively high scores in reported metrics. The quality of both classification tasks can be readily observed via confusion matrices given in Figure 2 and Figure 3. For example, the INF2PRES model classified 92% verbs that belong to the *-am* class accurately, and misclassified only 8%. In the PRES2INF model, the *-ati* verbs were

<sup>10</sup>[https://github.com/dseverdi/HR\\_verb\\_classification](https://github.com/dseverdi/HR_verb_classification)

Table 6:  
Classification performance  
with and without FastText  
character embeddings

	Accuracy	Micro- $F_1$	Macro- $F_1$	Weighted- $F_1$
INF2PRES	0.896	0.896	0.877	0.896
+ FastText	<b>0.947</b>	<b>0.947</b>	<b>0.936</b>	<b>0.947</b>
PRES2INF	0.931	0.931	0.829	0.925
+ FastText	<b>0.947</b>	<b>0.947</b>	<b>0.834</b>	<b>0.943</b>

Figure 2:  
INF2PRES classification

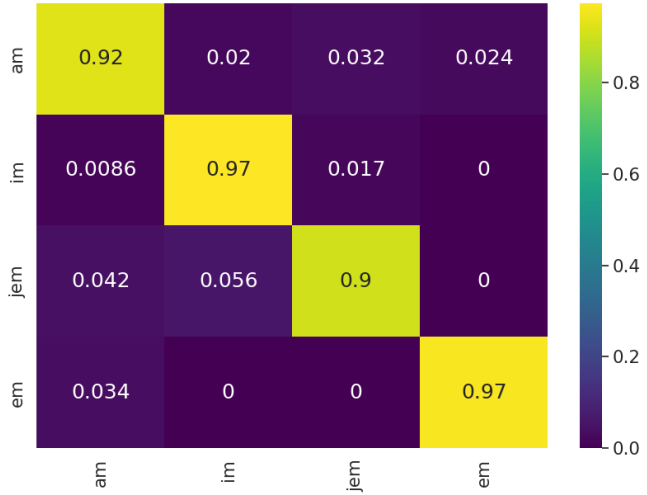
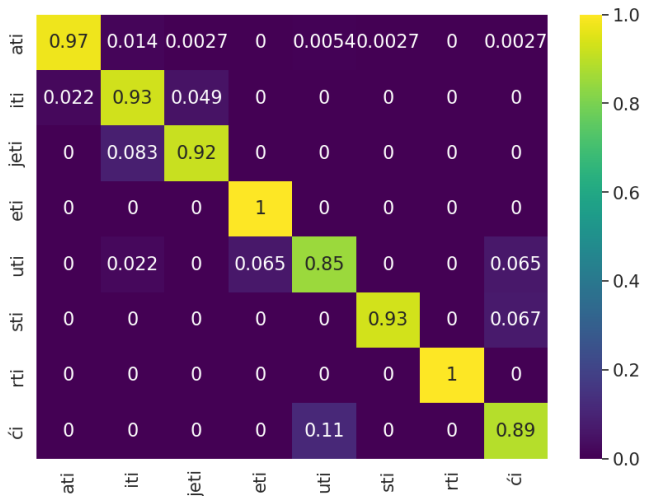


Figure 3:  
PRES2INF classification



classified properly in 97% cases, with only 3% of misclassifications. In a good classifier, diagonal values in confusion matrices should be as high as possible.

The interesting thing to see in our experiments are the CAMs for characters of verbs shown in Figure 4 and Figure 5.

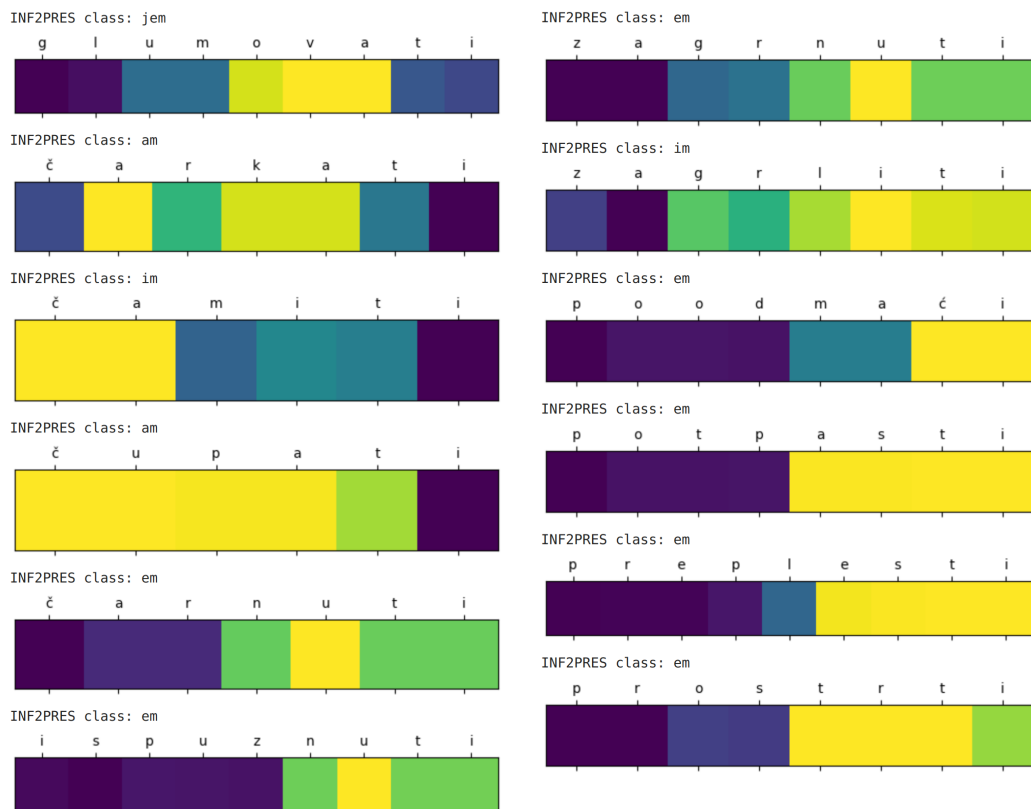


Figure 4: CAM filters for INF2PRES correctly predicted examples. Lighter colors indicate high contributing characters to classification

INF2PRES: Most CAM highlighted characters are usually suffixes of the verb with few exceptions. For example, verbs like *čarnuti* ‘to ignite’, *ispuznuti* ‘to slide off’, *zagnuti* ‘to cover’, *zagrliti* ‘to hold’, *podmaći* ‘to go off’, *potpasti* ‘to fall under’, and *prostrti* ‘to lay down’ follow this pattern. In some cases, infixes have more significance, as in the

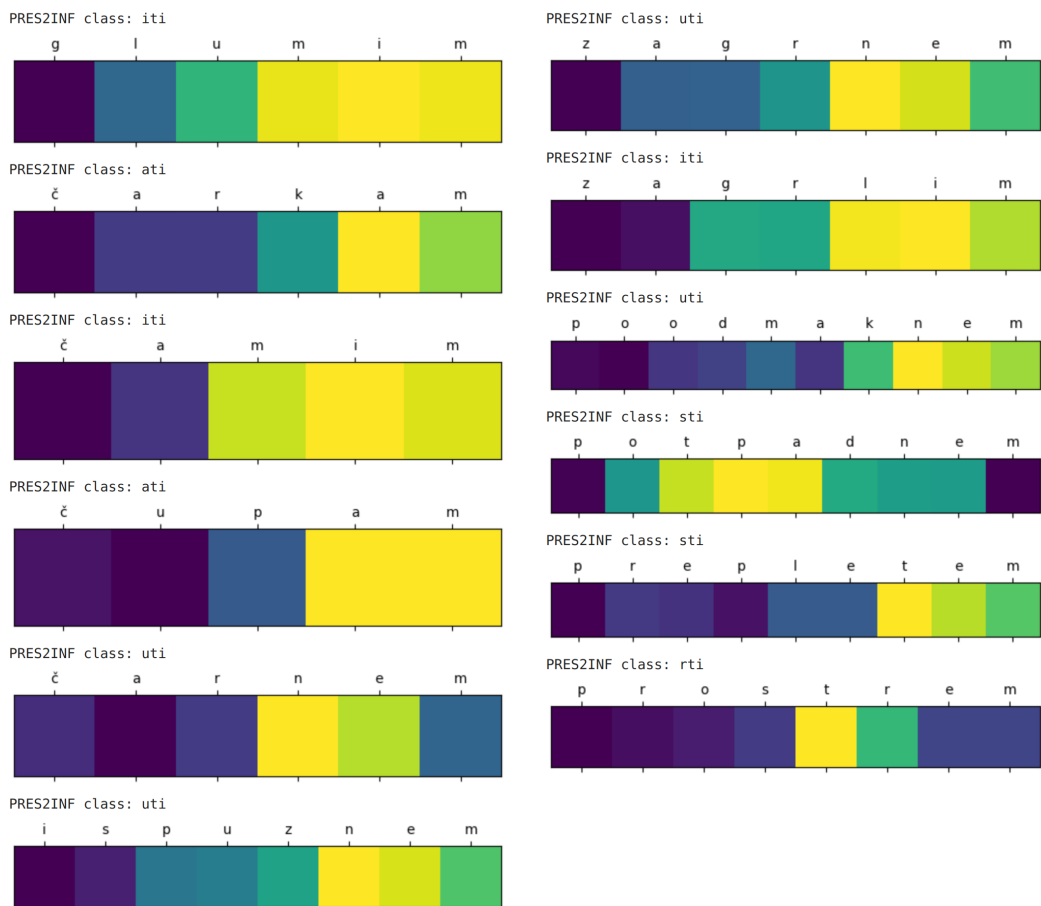


Figure 5: CAM filters for PRES2INF correctly predicted examples. Lighter colors indicate high contributing characters to classification

verb *glumovati* ‘to act’. For the verbs *čarkati* ‘to quarrel’, *čamiti* ‘to wait’, and *čupati* ‘to twitch’, it highlights initial phoneme clusters more than suffixes, as the *-ati* class is more ambiguous.

**PRES2INF:** In most cases, the suffix has a significant role in classification. The significance of the infix is given in *potpadnem* ‘I fall under’ and *prostrem* ‘I lay down’. For the latter, it is used to differentiate classes *-rti*, *-sti*. Note that, in both classification tasks, the model deals well with compound verbs (*odmaći*→*poodmaći*, *plesti*→*preplesti*, ...).



Unsurprisingly, the CAM filters of Figures 4 and 5 show chaotic patterns of what the important letters contribute to classifying verbs into classes. Sometimes the important letters are at the beginning, sometimes in the middle, and sometimes at the end of the word. It seems that the classifier is having a hard time finding real regularities. However, this is somewhat expected. If it were not the case, such classification would be relatively straightforward. For example, the verbs *glumovati* ‘to act’ and *gladovati* ‘to starve’ have the first person present form *glumujem* and *gladujem*, respectively. In contrast, the verb *glumatati* ‘to pretend’ has the first person present *glumim*. The model in this example is not prone to give importance to suffixes because, as a feature, they are not significant for classification. However, if the model attends more to the morphemes *-ova-* and *-at-* while considering that both verbs end with *-ati*, it can infer what the appropriate classes are.

### Ablation study

3.3

In this section, we consider the importance of specific architectural concepts of our model, namely:

- *importance of windowing*: The baseline model denoted as  $\text{FF}^w$  is a feed-forward neural network with two layers. The top layer is a softmax classifier. The purpose of the first layer is to find a mapping of aggregated information from the characters within a window of size  $w$  (concatenation of character vectors within a window). The second layer aims to find high-level features for the classification.
- *importance of convolutions*: a CNN model denoted  $\text{CNN}^{\{k_r\}}$  described in Section 2.1, with the list of 1D convolution sizes  $\{k_r\}$  with a total of  $l$  channels.

All models use ReLU as an activation function and are trained using cross-entropy loss using the ADAM optimizer. The metaparameters are set as in Table 5 and FastText pretrained character vectors are used. No model performance degradation due to class imbalance was observed using standard cross-entropy training compared to cross-entropy with class weights.

Table 7: INF2PRES models performance. Arrows indicate whether greater or lower is better

Model	Accuracy $\uparrow$	RMSE $\downarrow$	Micro-F1 $\uparrow$	Macro-F1 $\uparrow$	Weighted-F1 $\uparrow$
FF <sup>1</sup>	0.859	0.881	0.835	0.859	0.859
FF <sup>5</sup>	0.885	0.781	0.863	0.885	0.884
CNN <sup>{5}</sup>	0.939	0.565	0.939	0.928	0.939
CNN <sup>{1,2,3,5}</sup>	0.941	0.534	0.941	0.927	0.947

Table 8: PRES2INF models performance. Arrows indicate whether greater or lower is better

Model	Accuracy $\uparrow$	RMSE $\downarrow$	Micro-F1 $\uparrow$	Macro-F1 $\uparrow$	Weighted-F1 $\uparrow$
FF <sup>1</sup>	0.894	1.016	0.643	0.894	0.879
FF <sup>5</sup>	0.913	0.773	0.703	0.913	0.907
CNN <sup>{5}</sup>	0.946	0.529	0.946	0.783	0.943
CNN <sup>{1,2,3,5}</sup>	0.950	0.467	0.950	0.835	0.948

In both Table 7 and Table 8, one can observe that the addition of windows improves performance over the baseline model. The reason for this is that windows make it possible to capture the local context of characters. Moreover, the filtering of characters with only one convolution with  $l$  filters was beneficial for the model. We believe that multiple channels in convolution enabled the capture of several aspects of features for classification. The addition of several convolution sizes slightly improved the overall result.

### 3.4

#### *Experiments with SSF*

The SSF (Orešković et al. 2016)<sup>11</sup> contains a rule-based morphological generator (MG) for expanding its Croatian lexicon. It is written in Python and included in the SSF as a web service. The whole of SSF’s lexicon was processed initially by the MG, manually corrected and published as an online resource in the Linguistic Linked Open Data cloud (Orešković et al. 2018). The MG in general takes a lemma of

<sup>11</sup> SSF is publicly available at <http://ss-framework.com/?lang=en>.

	Accuracy	RMSE	Micro- $F_1$	Macro- $F_1$	Weighted- $F_1$
SSF INF2PRES	0.667	0.047	0.667	0.635	0.658

Table 9:  
SSF rule-based  
morphological parser  
performance on test set



Figure 6:  
SSF INF2PRES  
classification

the word and tries to find its proper inflections by applying specific grammatical rules. In the current state, it does not use any statistical information about words. Specifically for verbs, it takes an infinitive as an input and applies cascading rules that extract the root of the verb by subtracting known suffixes. Once the root is extracted, MG merges root and suffix for each verb form. For present tense, suffixes are: *-em*, *-im*, *-jem*, *-am*. After the root and suffix are merged, the MG applies sound changes to that newly formed word (e.g. sibilization, palatalization, iotation, etc.). Using a strictly rule-based approach, MG ends up with several equally probable paradigms (i.e. it applies at least one of the possible present tense suffixes). It is worth noting that it is still in development and using CAMs from INF2PRES model, and it can help us derive, to a certain extent, meaningful rules for better MG transduction. The current performance of the SSF for INF2PRES classification is given in Table 9 and Figure 6, if we choose only one paradigm (i.e. the first one). We do not apply it on PRES2INF because it is primarily designed for infinitive input.

Infinitiv			
	☐ <b>peglati</b> peglati		

Prezent			
	Odaberi	Odaberi	Odaberi
Jed 1	☐ <b>peglam</b> peglam	☐ <b>peglim</b> peglim	☐ <b>pegljem</b> pegljem
Jed 2	☐ <b>peglaš</b> peglaš	☐ <b>pegliš</b> pegliš	☐ <b>pegljem</b>
Jed 3	☐ <b>pegla</b> pegla	☐ <b>pegli</b> pegli	☐ <b>peglje</b> peglje
Množ 1	☐ <b>peglamo</b> peglamo	☐ <b>peglimo</b> peglimo	☐ <b>pegljemo</b> pegljemo
Množ 2	☐ <b>peglate</b> peglate	☐ <b>peglite</b> peglite	☐ <b>pegljete</b> pegljete
Množ 3	☐ <b>peglaju</b> peglaju	☐ <b>pegle</b> pegle	☐ <b>peglju</b> peglju


Infinitiv			
	☐ <b>putovati</b> putovati		

Prezent			
	Odaberi	Odaberi	Odaberi
Jed 1	☐ <b>putovam</b> putovam	☐ <b>putovim</b> putovim	☐ <b>putujem</b> putujem
Jed 2	☐ <b>putovaš</b> putovaš	☐ <b>putoviš</b> putoviš	☐ <b>putujem</b>
Jed 3	☐ <b>putova</b> putova	☐ <b>putovi</b> putovi	☐ <b>putuje</b> putuje
Množ 1	☐ <b>putovamo</b> putovamo	☐ <b>putovimo</b> putovimo	☐ <b>putujemo</b> putujemo
Množ 2	☐ <b>putovate</b> putovate	☐ <b>putovite</b> putovite	☐ <b>putujete</b> putujete
Množ 3	☐ <b>putovaju</b> putovaju	☐ <b>putove</b> putove	☐ <b>putuju</b> putuju

p e g l a t i



p u t o v a t i




Figure 7: This snippet of MG conjugation shows 3 possible inflections for the verbs *peglati* and *putovati* (only present is shown). Our classifier predicts which inflection is suitable: *peglati*→*peglam*, *putovati*→*putujem*. Note that both examples have the same suffix *-ati* but have different present stems

For example (see Figure 7), MG yields 3 groups for the verb *peglati* ‘to iron’ and our model picks the correct one:

- *pegl + am*
- *\*pegl + im*
- *\*pegl + jem*

CAM: attend to the stem boundary and thematic suffix *-a* and start of the final morpheme.

For the verb *putovati* ‘to travel’, MG produces also 3 possible conjugations (with *-ova* as a thematic suffix) and our model picks the correct one:

- *\*put + ov + am*
- *\*put + ov + im*
- *put + u + jem*

CAM: attend more to the thematic suffix *-ova* and start of the final morpheme.

In these examples, CAM shows the significance that the model assigns to each character with regard to the choice of present final morphemes. For future work, this can be helpful to define rules for MG to capture the proper inflection.

### *Experiments with SIGMORPHON 2023 baselines*

3.5

We compare our results with the baselines of SIGMORPHON 2023 Task 0 (Part 1), namely:

- `non-neural` model: a simple model that tries to align input/output examples during the training using Levenshtein distance and deduce appropriate prefix and suffix changing rules for given examples
- `neural` model: a Transformer based model applied for character level transduction from Wu *et al.* 2021.

Both models are implemented and publicly available.<sup>12</sup>

In our setup, they were trained and validated on INF2PRES datasets as generative models, i.e., they predict the proper inflected verb for the given infinitive. We treat the problem as a classification task.

Table 10 shows results for transducing Croatian verbs from infinitive (lemma) to first person present. For comparison, we also show our CNN model combined with MG for INF2PRES transduction. Although SIGMORHPON models achieve relatively worse results in our setup, they should be the first choice if datasets are large enough so that these models can learn general inflections (not constrained to verbs only). Our approach is more restricted, and it is useful if data is scarce and if rule-based systems are available (which is the case for the Croatian language).

---

<sup>12</sup><https://github.com/sigmorphon/2023InflectionST/tree/main/part1/baselines>

Table 10:  
SIGMORPHON 2023 baselines  
for Croatian verb transduction

Model	Accuracy
non-neural	0.8786
neural	0.8994
CNN + MG	0.9470

## CONCLUSION

In this paper, we provide an overview and motivation for the Croatian verb classification problem as a particular case of the Slavic inflection system. A neural network model with class activation mapping was applied as a supervised learning model on collected datasets. It is the initial step in applying present and infinitive stem classifiers in conjugating Croatian verbs. From this point on, one can apply rule-based transducers designed explicitly for the Croatian language (SSF by Orešković *et al.* 2018 would be one example) or apply some of the tools available on the market. If there is an abundance of data, one should resort to the established state-of-the-art models available via SIGMORPHON shared tasks.

Following the recent trends in natural language processing, the shift from rule-based and predictive models (supervised learning) to generative or unsupervised models becomes an interesting approach in inflectional morphology, especially for morphologically rich languages like Croatian. There are some promising results that encourage this pursuit, such as those in Şulea and Young 2019.

## REFERENCES

P. J. ANTONY, Hemant B. RAJ, B. S. SAHANA, Dimple Sonal ALVARES, and Aishwarya RAJ (2012), Morphological analyzer and generator for Tulu language: A novel approach, in Sabu M. THAMPI, El-Sayed EL-AFRY, and Javier AGUIAR, editors, *Proceedings of the International Conference on Advances in Computing, Communications and Informatics, ICACCI '12*, pp. 828–834, Association for Computing Machinery, New York, NY, USA, doi:10.1145/2345396.2345531.

Eugenija BARIĆ, Mijo LONČARIĆ, Dragica MALIĆ, Slavko PAVEŠIĆ, Mirko PETI, Vesna ZEČEVIĆ, and Maja ZNIKA (2005), *Hrvatska gramatika*, Školska knjiga, Zagreb, ISBN 9789530400108.

Cristina BARROS, Dimitra GKATZIA, and Elena LLORET (2017), Inflection generation for Spanish verbs using supervised learning, in Manaal FARUQUI, Hinrich SCHUETZE, Isabel TRANCOSO, and Yadollah YAGHOOBZADEH, editors, *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 136–141, Association for Computational Linguistics, Copenhagen, Denmark, doi:10.18653/v1/W17-4120.

Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN, and Tomas MIKOLOV (2017), Enriching word vectors with subword information, *Transactions of the Association for Computational Linguistics*, pp. 135–146, doi:10.1162/tacl\_a\_00051.

Marina DANILEVSKY, Kun QIAN, Ranit AHARONOV, Yannis KATSIS, Ban KAWAS, and Prithviraj SEN (2020), A survey of the state of explainable AI for natural language processing, in Kam-Fai WONG, Kevin KNIGHT, and Hua WU, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 447–459, Association for Computational Linguistics, Suzhou, China.

Liviu P. DINU, Vlad NICULAE, and Octavia-Maria SULEA (2012), Learning how to conjugate the Romanian verb. Rules for regular and partially irregular verbs, in Walter DAELEMANS, editor, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 524–528, Association for Computational Linguistics, Avignon, France.

Josef DOBROVSKÝ (1809), *Ausführliches Lehrgebäude der Böhmischen Sprache, zur gründlichen Erlernung derselben für Deutsche, zur vollkommenern Kenntniß für Böhmen*, J. Herrl, Prague, [https://books.google.fr/books?vid=U0M:39015036760190&redir\\_esc=y](https://books.google.fr/books?vid=U0M:39015036760190&redir_esc=y).

Greg DURRETT and John DENERO (2013), Supervised learning of complete morphological paradigms, in Lucy VANDERWENDE, Hal DAUMÉ III, and Katrin KIRCHHOFF, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1185–1195, Association for Computational Linguistics, Atlanta, Georgia, <https://aclanthology.org/N13-1138>.

Vishal GOYAL and Gurpreet Singh LEHAL (2008), Hindi morphological analyzer and generator, in Preeti R. BAJAJ, Amol Y. DESHMUKH, and Kailash D. JOSHI, editors, *2008 First International Conference on Emerging Trends in Engineering and Technology*, pp. 1156–1159, doi:10.1109/ICETET.2008.11.

Nizar HABASH and Owen RAMBOW (2006), MAGEAD: A morphological analyzer and generator for the Arabic dialects, in Nicoletta CALZOLARI, Claire

- CARDIE, and Pierre ISABELLE, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 681–688, Association for Computational Linguistics, Sydney, Australia, doi:10.3115/1220175.1220261.
- Diederik P. KINGMA and Jimmy BA (2015), ADAM: A method for stochastic optimization, in Yoshua BENGIO and Yann LECUN, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, doi:10.48550/arXiv.1412.6980.
- Mikhail KOROBV (2015), Morphological analyzer and generator for Russian and Ukrainian languages, in Mikhail Yu. KHACHAY, Natalia KONSTANTINOVA, Alexander PANCHENKO, Dmitry IGNATOV, and Valeri G. LABUNETS, editors, *Analysis of Images, Social Networks and Texts*, pp. 320–332, Springer International Publishing, Cham, doi:10.1007/978-3-319-26123-2\_31.
- Gichang LEE, Jaeyun JEONG, Seungwan SEO, CzangYeob KIM, and Pilsung KANG (2018), Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network, *Knowledge-Based Systems*, 152(C):70–82, ISSN 0950-7051, doi:10.1016/j.knosys.2018.04.006.
- Horace G. LUNT (2001), *Old Church Slavonic grammar*, Mouton de Gruyter, The Hague, doi:10.1515/9783110876888.
- Lew R. MICKLESEN (1974), The common Slavic verbal system, in Ladislav MATEJKA, Victor TERRAS, and Anna CIENCALA, editors, *Vol. 1 Linguistics and Poetics*, chapter American contributions to the Seventh International Congress of Slavists, August 21–27, 1973, pp. 241–274, De Gruyter Mouton, Berlin, Boston, ISBN 9783110873948, doi:10.1515/9783110873948-011.
- Milan MIHALJEVIĆ (2014), *Slavenska poredbena gramatika 2. dio: Morfologija, prozodija, slavenska pradomovina.*, Školska knjiga, Zagreb, ISBN 953-0-30225-8.
- Marko OREŠKOVIĆ, Sandra LOVRENČIĆ, and Mario ESSERT (2018), Croatian Network Lexicon within the Syntactic and Semantic Framework and LLOD Cloud, *International Journal of Lexicography*, 32(2):207–227, ISSN 0950-3846, doi:10.1093/ijl/ecy024.
- Marko OREŠKOVIĆ, Jakov TOPIĆ, and Mario ESSERT (2016), Croatian linguistic system modules overview, in George Meladze TINATIN MARGALITADZE, editor, *Proceedings of the 17th EURALEX International Congress*, pp. 280–283, Ivane Javakhishvili Tbilisi University Press, Tbilisi, Georgia, ISBN 978-9941-13-542-2.
- Adam PASZKE, Sam GROSS, Francisco MASSA, Adam LERER, James BRADBURY, Gregory CHANAN, Trevor KILLEEN, Zeming LIN, Natalia GIMELSHEIN, Luca ANTIGA, Alban DESMAISON, Andreas KOPF, Edward YANG, Zachary DEVITO, Martin RAISON, Alykhan TEJANI, Sasank CHILAMKURTHY, Benoit STEINER, Lu FANG, Junjie BAI, and Soumith CHINTALA (2019),



PyTorch: An imperative style, high-performance deep learning library, in Hanna M. WALLACH, Hugo LAROCHELLE, Alina BEYGELZIMER, Florence D'ALCHÉ-BUC, Edward A. FOX, and Roman GARNETT, editors, *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., doi:10.48550/arXiv.1912.01703.

Ida RAFFAELLI, Marko TADIĆ, Božo BEKAVAC, and Željko AGIĆ (2008), Building croatian wordnet, in Attila TÁNACS, Dóra CSENDES, Veronica VINCZE, Christiane FELLBAUM, and Piek VOSSEN, editors, *Proceedings of the 4th Global WordNet Conference (GWC 2008)*, pp. 349–359, Global WordNet Association, Szeged, Hungary, ISBN 978-963-482-854-9.

Josip SILIĆ and Ivo PRANJKOVIĆ (2005), *Gramatika hrvatskoga jezika*, Školska knjiga, Zagreb.

Octavia-Maria ȘULEA and Steve YOUNG (2019), Unsupervised inflection generation using neural language modelling, in Ignacio ROJAS, Gonzalo JOYA, and Andreu CATALA, editors, *Advances in Computational Intelligence*, pp. 668–678, Springer International Publishing, Cham, doi:10.48550/arXiv.1912.01156.

Krešimir ŠOJAT, Matea SREBAČIĆ, and Marko TADIĆ (2012), Derivational and semantic relations of Croatian verbs, *Journal of Language Modelling*, 0(1):111–142, ISSN 2299-8470, doi:10.15398/jlm.v0i1.34.

Richard Howard WICENTOWSKI (2002), *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*, Ph.D. thesis, The Johns Hopkins University, <https://www.cs.swarthmore.edu/~richardw/pubs/thesis.pdf>.

Shijie WU, Ryan COTTERELL, and Mans HULDEN (2021), Applying the transformer to character-level transduction, in Paola MERLO, Jorg TIEDEMANN, and Reut TSARFATY, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1901–1907, Association for Computational Linguistics, Online, doi:10.18653/v1/2021.eacl-main.163.

Andrea ZIELINSKI, Christian SIMON, and Tilman WITTL (2009), Morphisto: Service-oriented open source morphology for German, in Cerstin MAHLOW and Michael PIOTROWSKI, editors, *State of the Art in Computational Morphology*, pp. 64–75, Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-04131-0\_5.

*Domagoj Ševerdija*

ORCID 0000-0001-9501-1025  
dseverdi@mathos.hr

Department of Mathematics,  
University J. J. Strossmayer of Osijek  
Trg Ljudevita Gaja 6  
31 000 Osijek, Croatia

*Rebeka Čorić*

ORCID 0000-0002-2388-385X  
rcoric@mathos.hr

Department of Mathematics,  
University J. J. Strossmayer of Osijek  
Trg Ljudevita Gaja 6  
31 000 Osijek, Croatia

*Marko Orešković*

ORCID 0000-0002-3723-9256  
moreskovic@nsk.hr

National and University Library  
in Zagreb  
Hrvatske bratske zajednice 4,  
10000 Zagreb, Croatia

*Lucian Šošić*

ORCID 0000-0002-1523-493X  
luciansosic@gmail.com

Faculty of Humanities  
and Social Sciences in Split  
Poljička cesta 35,  
21000 Split, Croatia

Domagoj Ševerdija, Rebeka Čorić, Marko Orešković, and Lucian Šošić (2024),  
*Detecting inflectional patterns for Croatian verb stems using class activation  
mapping*, *Journal of Language Modelling*, 12(1):43–68

DOI <https://dx.doi.org/10.15398/jlm.v12i1.347>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

CC BY <http://creativecommons.org/licenses/by/4.0/>

# Control, inner topicalisation, and focus fronting in Mandarin Chinese: modelling in parallel constraint-based grammatical architecture

*Chit-Fung Lam*  
University of Manchester

## ABSTRACT

This paper proposes a formal analysis of two displacement phenomena in Mandarin Chinese, namely inner topicalisation and focus fronting, capturing their correlational relationships with control and complementation. It examines a range of relevant data, including corpus examples, to derive empirical generalisations. Acceptability-judgment tasks, followed by mixed-effects statistical models, were conducted to provide additional evidence. This paper presents a constraint-based lexicalist proposal that is couched in the framework of Lexical-Functional Grammar (LFG). The lexicon plays an important role in regulating the behaviour of complementation verbs as they participate in the displacement phenomena. Unlike previous analyses that cast inner topicalisation and focus fronting as restructuring phenomena, this lexicalist proposal does not rely on hypothesised clause-size differences. It captures the empirical properties more accurately and accounts for a wider range of empirical patterns. Adopting the formally explicit framework of LFG, this proposal uses constraints that have mathematical precision. The constraints are computationally implemented using the grammar engineering tool Xerox Linguistic Environment, safeguarding their precision.

*Keywords:*  
*control,*  
*complementation,*  
*inner*  
*topicalisation,*  
*focus fronting,*  
*long-distance*  
*dependency,*  
*restructuring,*  
*Chinese,*  
*Lexical-Functional*  
*Grammar,*  
*acceptability-*  
*judgment*  
*experiments,*  
*grammar*  
*engineering*

This paper centres on the syntax of two displacement phenomena in Mandarin Chinese, namely inner topicalisation and focus fronting (Ernst and Wang 1995; Grano 2015; Huang 2018; Paris 1998; Paul 2002, 2005, 2015; Shyu 1995), exploring their interaction with control and complementation.<sup>1</sup> Inner topicalisation, also known as “object preposing”, involves an object relation being displaced to a position between the subject and the verb without an additional marker.<sup>2</sup> (1a) is an example of inner topicalisation, where *gugong* ‘imperial palace’ is the preposed object. (1b) shows the canonical SVO word order without object preposing.<sup>3</sup>

- (1) a. women [gugong]            qu-guo le  
       1PL        imperial.palace go-PFV SFP  
       ‘We have been to the imperial place before.’  
       (Paul 2002, p. 697)

---

<sup>1</sup>The paper is based on part of the author’s PhD project (Lam 2023). It also contains some revised findings of inner topicalisation previously discussed by Lam (2022) in the *Proceedings of the LFG’22 Conference*. I am very grateful for the guidance of my PhD supervisors, Kersti Börjars and Eva Schultze-Berndt, and for their comments on various drafts. Many thanks to the audiences at the LFG22, SE-LFG31, and NACCL-34 conferences for their comments on early drafts. I also would like to thank all the participants of the acceptability-judgment tasks. I greatly appreciate Ziling Bai’s help in offering additional native-speaker judgments on the language data. Last but not least, I would like to thank the anonymous reviewers for their valuable feedback as well as the amazing editorial team at *JLM* for processing my manuscript. Any error is mine.

<sup>2</sup>In Chinese, inner topicalisation is distinguished from external topicalisation, where the preposed object appears before the subject. The constraints discussed in this paper are applicable to inner topicalisation but not to external topicalisation. For a comparison between inner topicalisation and external topicalisation, see, e.g., Paul 2002, 2015.

<sup>3</sup>The following are the abbreviations used in the morpheme-by-morpheme glosses of this paper: CL = classifier, COMP = complementiser, C.SELF = complex reflexive, DE = pre-nominal modification marker, EXP = experiential, PFV = perfective, PL = plural, PRT = particle, SELF = simplex reflexive, SFP = sentence-final particle, SG = singular.

- b. women qu-guo [gugong] le  
1PL go-PFV imperial.palace SFP  
'We have been to the imperial place before.'

As for focus fronting, this paper centres on the type involving the focus marker *lian* 'even', with the fronted *lian* 'even' constituent being an object relation.<sup>4</sup> In (2a), the *lian* 'even' constituent is the object of *renshi* 'know'. Note that a *lian* 'even' constituent cannot remain in situ (i.e., in the canonical object position), as shown by (2b).<sup>5</sup>

- (2) a. wo-de pengyou [lian ta] dou renshi  
1SG-DE friend even 3SG PRT know  
'My friends know even him.'  
(Paul 2002, p. 700)
- b. \*wo-de pengyou dou renshi [lian ta]  
1SG-DE friend PRT know even 3SG  
'My friends know even him.'

Intriguing patterns emerge in such structures. As observed by Ernst and Wang (1995), Qu (1995), Paul (2002, 2005, 2015), and others, the inner topic or focus-fronting phrase must remain inside the complement clause of a non-control complementation verb (e.g., *shuo* 'say'). In (3a), the displaced object *na-ben xiaoshuo* 'that novel' occupies the post-subject position in the complement clause. In (3b), moving the displaced object into the matrix clause is ungrammatical.

- (3) a. wangwu shuo lisi [na-ben xiaoshuo] du-wan-le  
Wangwu say Lisi that-CL novel read-finish-PFV  
'Wangwu said that Lisi finished reading that novel.'
- b. \*wangwu [na-ben xiaoshuo] shuo lisi du-wan-le  
Wangwu that-CL novel say Lisi read-finish-PFV  
'Wangwu said that Lisi finished reading that novel.'  
(Ernst and Wang 1995, p. 244)

---

<sup>4</sup> Another focus-fronting construction discussed in the literature involves fronting a *shenme* 'what' constituent.

<sup>5</sup> In focus-fronting, the particle *dou* is usually needed to make the construction well-formed. Although some references e.g., Huang *et al.* (2009) translate *dou* as 'all', it does not preserve much (if any) of the meaning of "all".

In contrast, for control verbs (e.g., *shefa* ‘try’), it has been reported that the inner topic or focused phrase occupies a post-subject position in the matrix clause (Grano 2015; Huang 2018). In (4), the displaced object *zhe-pian baogao* ‘this report’ appears after the matrix subject *wo* ‘I’, crossing the control verb *shefa* ‘try’.

- (4) *wo* [*zhe-pian baogao*] *hui shefa jinkuai xie-wan*  
1SG this-CL report will try soon write-finish  
‘I will try to finish even this report soon.’  
(Huang 2018, p. 351)

The displacement phenomena seem to correlate with the (non-)control status of the complementation verb. Further discussion about different types of control will be provided with regard to how they correlate with the displacement phenomena.

This paper aims to model the intricate relationships among control, inner topicalisation, and focus fronting. The formal analysis will be couched in the framework of Lexical-Functional Grammar (LFG; Bresnan 1982; Bresnan *et al.* 2016; Dalrymple *et al.* 2019), which is a formally explicit grammatical theory that uses constraints of mathematical precision. This approach provides a fresh analytical perspective, as most previous studies have been conducted within derivational frameworks (Principles & Parameters; Minimalism). The paper offers detailed empirical data on how the displacement phenomena interact with control and complementation, which can be valuable for researchers of different theoretical orientations.<sup>6</sup>

The paper is organised as follows. Section 2 introduces three classes of complementation verbs – exhaustive-control, partial-control, and non-control – which are relevant to the issues at hand. Section 3 critically reviews a Minimalist proposal, which approaches inner topicalisation and focus fronting as restructuring phenomena. Section 4 presents the relevant empirical patterns. It also reports the results of five acceptability-judgment tasks (AJTs) to provide additional evidence. Section 5 offers a pre-theoretical explanation for the empirical generalisations. Section 6 articulates the LFG grammati-

---

<sup>6</sup>This paper focuses on syntactic constraints. For a discussion regarding the information-structural properties of inner topicalisation and focus fronting, see, e.g., Ernst and Wang 1995, Paul 2002, Shyu 1995.

cal architecture as background information. Section 7 proposes an LFG formal analysis to capture the correlational relationships among control, inner topicalisation, and focus fronting. Section 8 brings in computational testing for the constraints in the formal analysis, drawing on LFG’s computational rigour. Section 9 concludes the paper.

EXHAUSTIVE-CONTROL  
VS PARTIAL-CONTROL  
VS NON-CONTROL VERBS

2

This paper centres on three classes of complementation verbs in the displacement phenomena: exhaustive-control vs partial control vs non-control verbs. The differences between exhaustive- and partial-control verbs are discussed in the general literature (e.g., Grano 2015; Haug 2013; Landau 2000, 2013). Crucially, an exhaustive-control verb requires strict identity between the controller and controllee, while a partial-control verb allows the entity denoted by the controller to be a subset of the entities denoted by the controllee. To differentiate between them, we use the “collective-word diagnostic”, which involves a semantically singular controller and a semantically plural controllee with a collective word (e.g., *yiqi* ‘together’, *jihe* ‘gather’) in the complement clause. (5) and (6) illustrate the diagnostic. The controller is the matrix subject *Xiaoming* and the controllee is the embedded subject (notated as “ $\emptyset$ ”). The results suggest that while *shefa* ‘try’, *deyi* ‘manage’, and *jinli* ‘endeavour’ are exhaustive-control verbs, *dasuan* ‘intend’, *xiangyao* ‘want’, and *jueding* ‘decide’ license partial control.

- (5) *xiaoming*<sub>i</sub> *shefa/deyi/jinli*  $\emptyset_{i/*j}$  #(gen pengyou) ba  
*Xiaoming* try/manage/endeavour  $\emptyset$  with friend eight  
*dianzhong* *jihe*  
 o'clock gather  
 ‘Xiaoming tries/manage/endeavour to gather #(with friends) at eight o'clock.’

- (6) xiaoming<sub>i</sub> dasuan/xiangyao/jueding  $\emptyset_{i+/*j}$  ba dianzhong  
 Xiaoming intend/want/decide  $\emptyset$  eight o'clock  
 jihe  
 gather  
 'Xiaoming intends/wants/decides to gather at eight o'clock.'

Note that outside the collective-word context, partial-control verbs allow complete coreference between the controller and controllee.

Chinese is a discourse pro-drop language (Huang 1984, 1989), allowing unexpressed subjects and objects. In a non-control complementation construction, when the embedded subject is unexpressed, the non-control verb (e.g., *shuo* 'say', *xiangxin* 'believe', *renwei* 'think') does not place coreferential constraints on it. The unexpressed embedded subject can refer to the matrix subject or another discourse-salient entity in a way similar to its pronominal counterpart, as shown in (7).

- (7) xiaoming<sub>i</sub> shuo/xiangxin/renwei  $\{\emptyset_{i/j} \mid ta_{i/j}\}$  jian-guo  
 Xiaoming say/believe/think  $\{\emptyset \mid 3SG\}$  see-EXP  
 zhangsan le  
 Zhangsan SFP  
 'Xiaoming says/believes/thinks (he) has seen Zhangsan.'

Section 4 onwards will demonstrate correlational relationships between these classes of verbs and their patterns in inner topicalisation and focus fronting.

### 3 AGAINST RESTRUCTURING APPROACHES TO INNER TOPICALISATION AND FOCUS FRONTING

In recent years, there has been a trend in the Minimalist tradition to understand inner topicalisation and focus fronting as restructuring (Grano 2015; Huang 2018), explaining the contrast between (3) and (4) based on clause-size differences. Restructuring is, in



essence, a clause-size-reduction phenomenon (Aissen and Perlmutter 1976; Cinque 2006; Rizzi 1978; Wurmbrand 2001, 2004, 2015). In the derivational tradition, while a control construction is typically characterised as a bi-clausal configuration where the complement clause projects up to CP (or at least TP), a subset of control verbs is said to select for a size-reduced embedded structure (e.g., non-clausal vP). Thus, the construction is said to display behaviour typically attested in a mono-clausal configuration. Several claims have been made regarding inner topicalisation and focus fronting based on restructuring. It has been claimed that whether the displaced object can “cross” the complementation verb is contingent on the size of the embedded complement. Assuming movement, it is posited that a control verb restructures its embedded complement into a non-clausal structure (Grano 2015) or a reduced clausal structure (Huang 2018) such that the displaced object moves across the boundary between the matrix clause and embedded complement, forming (4). On the contrary, a non-control verb forms a bi-clausal configuration with its embedded complement projecting up to a clausal domain, blocking any further movement of the displaced phrase; thus, the displacement is only viable within the embedded clause (Grano 2015; Huang 2018), explaining the patterns in (3).

The above claims are instantiated in Huang’s (2018) formal analysis of inner topicalisation, displayed in (8). In his formal system, InnerTopP is a projection in the “operator” domain (comparable to CP in the general literature), signalling a full-fledged clausal structure. After arriving at the InnerTopP position, an inner topic “freezes” due to some feature-checking mechanism. (8a) models inner topicalisation in a control construction. The embedded complement is restructured as a non-clausal vP. Without the CP domain (InnerTopP projection) in the embedded complement, the inner topic undergoes multiple movements, crossing the control verb and arriving at a post-matrix-subject position to satisfy some theory-internal feature-checking mechanism. (8b) models the movement of an inner topic in a non-control construction. Since a non-control construction lacks clausal restructuring, the CP domain (InnerTopP projection) is found in the embedded complement, stopping the inner topic from moving further upward.



plement of a control verb is restructured to a non-clausal structure.<sup>7</sup>

We offer one more empirical test – a complex reflexive binding diagnostic – to challenge the claim of restructuring. This diagnostic is based on the observation that the Mandarin complex reflexive *taziji* needs to be locally bound by a subject relation (Charnavel *et al.* 2017; Huang *et al.* 2009; Lam 2021). Part of its binding condition is stated in (10) (see Lam 2021 for further details):

- (10) When the complex reflexive *taziji* takes on a non-subject grammatical relation, *taziji* must be locally bound by the subject of the same verb which selects for *taziji*.

The diagnostic is applied to (11):

- (11) a. xiaoming [(lian) na-fen liwu] (dou) shefa (zai zuihou  
Xiaoming even that-CL gift PRT try at last  
guantou) song gei taziji  
moment give to C.SELF  
'Xiaoming tries to, at the last moment, give (even) that gift  
to himself.'

---

<sup>7</sup>Although Huang (2018) noticed the availability of *shuo* after control and non-control verbs, he treats it as a non-complementiser functional head in the inflectional domain. His treatment thus stands in contrast to Chappell's (2008) typological investigation on Chinese languages. However, as admitted by Huang (2018, p. 370) himself, his treatment of *shuo* has a few unresolved issues. Besides having to leave the exact functional category of *shuo* undetermined, he also needs to go against the cross-linguistic observation that SAY verbs (*verbal dicendi*) grammaticalise into complementisers (see, e.g., Chappell 2008) as well as to address a few distributional issues related to the fronting of a constituent before *shuo*. Overall, Huang (2018) does not provide independent empirical evidence to substantiate the claim that the embedded complement of a control verb is smaller than that of a non-control verb in cases of inner topicalisation or focus fronting. As the suggested difference in clause size is used to explain their distinct behaviour in inner topicalisation or focus fronting, attempts to posit this behaviour as evidence for the difference in clause size would amount to circular reasoning.

- b. xiaoming xiangxin (ta) [(lian) na-fen liwu] (dou) hui  
 Xiaoming believe 3SG even that-CL gift PRT will  
 (zai zuihou guantou) song gei taziji  
 at last moment give to C.SELF  
 ‘Xiaoming believes that (he) will, at the last moment, give  
 (even) that gift to himself.’

Being the oblique object of the embedded verb, *taziji* needs to be locally bound by its subject. The fact that (11a) and (11b) are well-formed suggests that there must be an (unexpressed) subject for the embedded verb *gei* ‘give’, serving as the antecedent of *taziji* in order to satisfy its binding requirement. The presence of an embedded subject suggests clausal embedding (see Butt 2014). That means both control and non-control constructions in (11) are bi-clausal, contrary to the claim that a control construction is restructured to be mono-clausal in inner topicalisation and focus fronting.

In LFG, clausehood is a multi-level concept (see, e.g., Butt 2014). The *shuo*-complementiser diagnostic signals clausehood at the phrase-structural level (c-structure), whereas the binding diagnostic reveals clausehood at the functional level (f-structure). More information about the two levels will be discussed in Section 6. Together, the diagnostics suggest that control and non-control constructions are bi-clausal at both phrase-structural and functional levels in inner topicalisation and focus fronting. Because there is no independent syntactic evidence to support clause-size differences, Huang’s (2018) restructuring analysis is empirically unfavourable.

Another shortcoming of Huang’s (2018) restructuring approach is that not all control verbs demonstrate the obligatory extraction pattern of (8a). For example, it is acceptable for the displaced object of a *dasuan* ‘intend’ construction to appear either at the post-matrix-subject position or inside the complement clause, as exemplified by (12):

- (12) a. xiaoming [zhe-xiang gongzuo] dasuan yao yiqi  
 Xiaoming this-CL task intend will together  
 wancheng  
 finish  
 ‘Xiaoming intends to finish this task together.’

- b. xiaoming dasuan [zhe-xiang gongzuo] yao yiqi  
Xiaoming intend this-CL task will together  
wancheng  
finish  
'Xiaoming intends to finish this task together.'

Although several studies (e.g., Hu *et al.* 2001, p. 1142; Huang 2018, p. 364; Zhang 2016, p. 291) have noticed the pattern of (12b), Huang (2018, p. 364) treats it as a (non-standard) variant arising from interspeaker variation. However, the recurrence of this pattern in different studies leads one to doubt whether this is truly the best treatment for the pattern. In fact, a crucial difference between (11) and (12) lies in the divergent control properties of *shefa* 'try' and *dasuan* 'intend' – the former an exhaustive-control verb while the latter a partial-control one. In other words, whether the displaced object can remain inside the complement clause correlates with the complementation verb's control behaviour. To the best of our knowledge, there is no existing study providing a formal mechanism to model such correlations.

Based on the above discussion, a movement-based restructuring approach to inner topicalisation and focus fronting is unsatisfactory. This paper will devise an alternative formal mechanism. Before that, the forthcoming section will clarify the empirical landscape of the two displacement phenomena in relation to control and complementation.

## EMPIRICAL GENERALISATIONS

4

This section presents five empirical generalisations regarding inner topicalisation, focus fronting, control, and complementation by examining qualitative data. Patterns A to D concern complementation constructions without a matrix object, while Pattern E pertains to object-control constructions. The five patterns were tested in acceptability-judgment tasks (AJTs) using a subset of the complementation verbs to provide additional quantitative evidence to supplement the qualitative discussion. Section 5 will provide some pre-theoretical insights into why exhaustive, partial, and non-control verbs behave in the ways described below.

4.1

*Pattern A: Exhaustive subject control  
and inner topicalisation/focus fronting*

If a complementation verb licenses exhaustive subject control, the displaced object must appear in the matrix clause, crossing the complementation verb. This pattern corroborates the judgments of Grano (2015) and Huang (2018). (13) illustrates this pattern with the exhaustive subject-control verbs *shefa* ‘try’, *xiangbanfu* ‘strive’, *changshi* ‘attempt’, *jujue* ‘refuse’, *deyi* ‘manage’ and *jinli* ‘endeavour’.<sup>8</sup>

- (13) a. xiaoming [(lian) zhe-jian shiqing] (dou) shefa/  
Xiaoming even this-CL matter PRT try/  
xiangbanfu/changshi/jujue/deyi/jinli duzi  
strive/attempt/refuse/manage/endeavour alone  
chuli  
handle  
‘Xiaoming tries/strives/attempts/refuses/manages/endeavours to handle (even) this matter alone.’
- b. \*xiaoming shefa/xiangbanfu/changshi/jujue/deyi/jinli  
Xiaoming try/strive/attempt/refuse/manage/endeavour  
[(lian) zhe-jian shiqing] (dou) duzi chuli  
even this-CL matter PRT alone handle  
‘Xiaoming tries/strives/attempts/refuses/manages/endeavours to handle this matter alone.’

4.2

*Pattern B: Partial subject control  
and inner topicalisation/focus fronting*

Partial subject-control verbs (e.g., *dasuan* ‘intend’, *zhunbei* ‘prepare’, *xiang(yao)* ‘want’, *jueding* ‘decide’, *kewang* ‘desire’, *zhiyi* ‘insist’, and *gan* ‘dare’) allow the displaced phrase to either cross the complemen-

<sup>8</sup>To see whether there are corpus examples that contradict the reported judgment here, we conducted corpus searches using the large-scale zhTenTen17 corpus via Sketch Engine <https://www.sketchengine.eu/zhtenten-chinese-corpus/>. Although there is no available keyword for inner topicalisation, we used the focus marker *lian* ‘even’ to construct CQL queries for the focus fronting of these exhaustive-control verbs. We tested the sequence of [exhaustive-control verb] + [*lian* ‘even’] and did not find any valid examples. On the other hand, we did find examples of [*lian* ‘even’] + NP + DOU + [exhaustive-control verb].

tation verb or remain inside the complement clause. (14) contains constructed examples. As discussed previously, the pattern of having the displaced object remaining inside the complement clause is not predicted by Huang's (2018) theoretical machinery. Data from the zhTenTen17 corpus (Jakubíček *et al.* 2013) and Google search results are provided below to support the acceptability of this pattern.<sup>9</sup>

- (14) a. xiaoming [(lian) na-ge difang] (dou) dasuan/  
Xiaoming even that-CL place PRT intend/  
zhunbei/xiangyao/jueding/kewang/zhiyi mingtian  
prepare/want/decide/desire/insist tomorrow  
(yao) yiqi qu  
will together visit  
'Xiaoming intends/prepares/wants/decides/insists to visit  
(even) that place tomorrow together.'
- b. xiaoming dasuan/zhunbei/xiangyao/juejing/kewang/  
Xiaoming intend/prepare/want/decide/desire/  
zhiyi [(lian) na-ge difang] mingtian (dou) yao  
insist even that-CL place tomorrow PRT will  
yiqi qu  
together visit  
'Xiaoming intends/prepares/wants/decides/insists to visit  
(even) that place tomorrow together.'
- (15) wo zhunbei jinhou [zhe-lei shu] duo kan yidian  
1SG prepare from.now this-kind book more read more  
'I prepare to read more of this kind of book from now on.'  
(Hu *et al.* 2001, p. 364)
- (16) pingguo shenzhi xiang [lian zuihou yi-ge shiti anjian]  
Apple even want even last one-CL physical button  
dou yao qudiao  
PRT will get.rid  
'Apple wanted to get rid of even the last physical button.'  
(zhTenTen17 corpus)

---

<sup>9</sup>The corpus data centre on focus fronting, as the focus marker *lian* 'even' lends itself to CQL queries; there is no similar keyword for inner topicalisation.

- (17) yamaxun jue ding [lian zhe-ge liwai] dou buzai  
Amazon decide even this-CL exception PRT no.longer  
baoliu  
keep  
'Amazon decided not to keep even this exception.'  
(zhTenTen17 corpus)
- (18) duifang zhiyi [lian yunfei] dou buyao wo chu  
other.party insist even shipping.fee PRT need.not 1SG pay  
'The other party insisted on not needing me to pay for the ship-  
ping fee.'  
(A Weibo post)<sup>10</sup>
- (19) ni jingran gan [lian ni shifu-de hua] dou bu  
you how.come dare even you master-DE word PRT not  
zuncong  
obey  
'How dare you do not obey even your master's words?'  
(zhTenTen17 corpus)

4.3

*Pattern C: Subject expression of partial control  
and inner topicalisation/focus fronting*

While partial-control verbs (e.g., *dasuan* 'intend', *zhunbei* 'prepare') usually require their embedded subject to be unexpressed, some verbs such as *jue ding* 'decide' and *kewang* 'desire' allow it to be optionally expressed. When the embedded subject is expressed as an overt pronoun, its reference follows its binding condition, unlike its unexpressed counterpart, whose reference is constrained to include the matrix subject. This observation is exemplified in (20), (21), and (22).<sup>11</sup>

<sup>10</sup> <https://weibo.com/1540060353/M2b7r7YOg>. Accessed on 10 Jan 2023.

<sup>11</sup> The co-indexation in (21) and (22) was added based on the contextual information of the corpus examples.



- (20) xiaoming<sub>i</sub> jueding/kewang { $\emptyset_{i+/*j}$  | tamen<sub>i+/\*j</sub>} mingtian  
 Xiaoming decide/desire { $\emptyset$  | 3PL} tomorrow  
 yiqi wancheng zhe-xiang gongzuo  
 together finish this-CL task  
 ‘Xiaoming decides/desires to finish this task together tomorrow.’/ ‘Xiaoming decides/desires that they will finish this task together tomorrow.’
- (21) shengwei<sub>i</sub> jueding ta<sub>j</sub> dao weinan ren  
 provincial.committee decide 3SG go Weinan serve  
 shiwei shuji  
 municipal.committee secretary  
 ‘The provincial party committee decided that he should go to Weinan to serve as the secretary of the municipal party committee.’  
 (zhTenTen17 corpus)
- (22) dang ta<sub>i</sub> jueding ta<sub>i</sub> xiang hui zhengfu gongzuo shi,  
 when he decide he want return government work time  
 men dou changkai-zhe  
 door all open-DUR  
 ‘When he decides that he wants to return to work in the government, the door will be open.’  
 (zhTenTen17 corpus)

This kind of partial-control verb is subject to an additional constraint. If the displaced phrase crosses the complementation verb, its embedded subject must be unexpressed. On the other hand, if the displaced phrase remains inside the complement clause, it is acceptable for the embedded subject to be either overt or unexpressed. This is illustrated in (23).

- (23) a. xiaoming [(lian) zhe-xiang gongzuo] (dou)  
 Xiaoming even this-CL task PRT  
 jueding/kewang (\*tamen) dei mingtian yiqi  
 decide/desire they should tomorrow together  
 wancheng  
 finish  
 ‘Xiaoming decides/desires to finish (even) this task together tomorrow.’

- b. xiaoming jueding/kewang (tamen) [(lian) zhe-xiang  
 Xiaoming decide/desire they even this-CL  
 gongzuo] (dou) dei mingtian yiqi wancheng  
 task PRT should tomorrow together finish  
 ‘Xiaoming decides/desires to finish (even) this task together  
 tomorrow.’

Although there is a difference between the sentence pair in (23) with regard to embedded-subject expression, we are not aware of any existing study documenting this observation.

4.4 *Pattern D: Non-control complementation  
 and inner topicalisation/focus fronting*

Non-control complementation verbs require their displaced phrase to reside in the complement clause, regardless of whether the embedded subject is overt or unexpressed (i.e., discourse pro-drop). This judgment has been reported in a number of studies (e.g., Ernst and Wang 1995; Grano 2015; Huang 2018; Paul 2002, 2005, 2015). (24) contains relevant examples with the non-control verbs *shuo* ‘say’, *xiangxin* ‘believe’, *renwei* ‘think’, *xiwang* ‘hope’, and *guji* ‘predict’.

- (24) a. \*xiaoming [(lian) zhe-ben shu] (dou) shuo/xiangxin/  
 Xiaoming even this-CL book PRT say/believe/  
 renwei/xiwang/guji { $\emptyset_{i/j}$  |  $ta_{i/j}$ } hui jinkuai  
 think/hope/predict { $\emptyset$  | 3SG} will soon  
 wancheng  
 complete  
 ‘Xiaoming says/believes/thinks/hopes/predicts he will  
 complete (even) this book soon.’
- b. xiaoming shuo/xiangxin/renwei/xiwang/guji [(lian)  
 Xiaoming say/believe/think/hope/predict even  
 zhe-ben shu] { $\emptyset_{i/j}$  |  $ta_{i/j}$ } (dou) hui jinkuai  
 this-CL book { $\emptyset$  | 3SG} PRT will soon  
 wancheng  
 complete  
 ‘Xiaoming says/believes/thinks/hopes/predicts he will  
 complete (even) this book soon.’

*Pattern E: Object control  
and inner topicalisation/focus fronting*

Patterns A–D apply to complementation verbs which do not select for an object, while Pattern E pertains to object-control verbs. For numerous object-control constructions, regardless of whether the verb licenses exhaustive control or partial control, it is not possible for the displaced phrase to cross the object controller and the phrase must remain inside the complement clause.<sup>12</sup> This pattern is exemplified in (25) and (26), which are constructed examples of inner topicalisation.

- (25) a. \*xiaoming [zhe-pian yanjiu baogao] pizhun/  
 Xiaoming this-CL research report permit/  
 quan/shuifu/guli/jiao/bi  
 try.to.persuade/persuade/encourage/ask/force  
 zhangsan tiqian san tian tijiao  
 Zhangsan in.advance three day submit  
 ‘Xiaoming permits/tries to persuade/persuades/ encour-  
 ages/asks/forces Zhangsan to submit this research report  
 three days in advance.’
- b. xiaoming pizhun/quan/shuifu/guli/  
 Xiaoming permit/try.to.persuade/persuade/encourage/  
 jiao/bi zhangsan [zhe-pian yanjiu baogao]  
 ask/force Zhangsan this-CL research report  
 tiqian san tian tijiao  
 in.advance three day submit  
 ‘Xiaoming permits/tries to persuade/persuades/ encour-  
 ages/asks/forces Zhangsan to submit this research report  
 three days in advance.’

---

<sup>12</sup>We have noted that object-raising verbs (e.g., *xiangyao* ‘want’ and *rang* ‘let’) as well as certain object-control verbs (e.g., *pai* ‘send’, *yaoqing* ‘invite’) appear to allow the displaced phrase to be positioned in the matrix clause. See Paul 2002 for some data regarding *rang* ‘let’ and *pai* ‘send’. Although we leave the explanation for future research, because this paper adopts a lexicalist approach to inner topicalisation and focus fronting, it is still feasible to independently formulate the relevant constraints for these individual verbs in their lexical entries to license their distinctive displacement behaviour (see Section 7).

- (26) a. \*xiaoming [zhe-pian yanjiu baogao] yuanliang/guai/  
 Xiaoming this-CL research report forgive/blame/  
 jinzhi zhangsan chichi bu tijiao  
 forbid Zhangsan delay not submit  
 ‘Xiaoming forgives/blames/forbids Zhangsan for/from de-  
 laying submitting this research report.’
- b. xiaoming yuanliang/guai/jinzhi zhangsan [zhe-pian  
 Xiaoming forgive/blame/forbid Zhangsan this-CL  
 yanjiu baogao] chichi bu tijiao  
 research report delay not submit  
 ‘Xiaoming forgives/blames/forbids Zhangsan for/from de-  
 laying submitting this research report.’

Examples (27)–(30) are corpus examples of focus fronting, demonstrating the acceptability of having the displaced phrase inside the complement clause.<sup>13</sup> Among these exemplified object-control verbs, *pizhun* ‘permit’, *jinzhi* ‘forbid’, *yuanliang* ‘forgive’, and *guai* ‘blame’ exhibit exhaustive control; whereas *quan* ‘try to persuade’, *shuifu* ‘persuade’, *guli* ‘encourage’, *jiao* ‘ask’, and *bi* ‘force’ exhibit partial control.<sup>14</sup>

- (27) tongcunren dou quan ta [lian shishou]  
 fellow.villagers all try.to.persuade 3SG even dead.body  
 dou bu bi yanmai  
 PRT not need bury  
 ‘The fellow villagers all tried to persuade him not to bury even  
 the dead body.’  
 (zhTenTen17 corpus)

- (28) nimen... bi wo [lian wo ge] dou bu qu jiu  
 2PL force 1SG even 1SG brother PRT not go save  
 ‘You all forced me not to go to save even my brother.’  
 (zhTenTen17 corpus)

<sup>13</sup>We also tried to look for counterexamples in the zhTenTen17 corpus with the displaced phrase appearing in the matrix clause for these object-control verbs, but we were not able to find relevant examples.

<sup>14</sup>In the general literature, control verbs such as *yuanliang* ‘forgive’ and *guai* ‘blame’ are semantically classified as factive verbs. See Landau 2000, pp. 45–46 for some cross-linguistic examples of factive verbs.

- (29) ta zhouwei-de ren... guai ta [lian yi-ge ren]  
3SG around-DE people blame 3SG even one-CL person  
dou shoushi buliao  
PRT defeat not.able.to  
'The people around him blamed him for not being able to defeat  
even one person.'  
(zhTenTen17 corpus)
- (30) qing yuanliang wo [lian mingzi] dou jibuzhu  
please forgive 1SG even name PRT cannot.remember  
'Please forgive me for not remembering even the name.'  
(zhTenTen17 corpus)

*Additional evidence from acceptability-judgment tasks*

4.6

The above section discussed five empirical generalisations (Patterns A–E). Besides cross-checking our reported judgments with corpus data, we also conducted five acceptability-judgment tasks (AJTs) on a subset of the complementation verbs.

*Design of acceptability-judgment tasks*

4.6.1

Each AJT tested one of the five generalisations. Each AJT adopted a  $2 \times 2$  factorial design, generating 4 conditions, each of which had 4 lexicalisations. Thus, there were 16 ( $= 4 \times 4$ ) test sentences for each task and, in total, 80 ( $= 5 \times 16$ ) test sentences across the five AJTs. The test sentences were distributed across eight lists using a Latin square design for counterbalancing. Lists 1–4 contained sentences for Tasks 1, 4, and 5. Lists 5–8 contained sentences for Tasks 2, 3, and 5. Each participant received one list, containing 4 test sentences for each task ( $= 12$  test sentences in total) and 13 fillers. No sentences in a list were variants of each other. The fillers were sentences of comparable syntactic complexity, displaying different degrees of acceptability. Among the fillers are constructions which should be highly acceptable and those which should be highly unacceptable. These “gold-standard” fillers were established based on a pilot run with other speakers beforehand. These fillers helped spot invalid responses to be

Table 1:  
2 × 2 factorial  
design of Task 1  
(Exhaustive  
Control), Task 2  
(Partial Control),  
and Task 4  
(Non-control)

	Crossing $V_m$	Not crossing $V_m$
Focus fronting	Crossing $V_m$ + focus fronting (Condition A)	Not crossing $V_m$ + focus fronting (Condition B)
Inner topicalisation	Crossing $V_m$ + inner topicalisation (Condition C)	Not crossing $V_m$ + inner topicalisation (Condition D)

discarded during data analysis.<sup>15</sup> All the sentences were randomised by Qualtrics, which was the survey tool used to distribute the AJTs.

Task 1 tested the generalisation that if a complementation verb licenses exhaustive control, the displaced phrase must precede the complementation verb (Pattern A). Task 2 tested the generalisation that for a partial-control verb, the displaced phrase can either precede the complementation verb or remain in the complement clause (Pattern B). Task 4 tested the generalisation that for a non-control complementation verb, the displaced phrase must remain in the complement clause (Pattern D). Table 1 presents the four testing conditions in each of the above-mentioned AJTs (Tasks 1, 2, and 4), with the displacement phenomena and positions of the displaced phrase as the independent variables. “ $V_m$ ” stands for complementation verb.

The four conditions are exemplified in Appendix A. The conditions for Task 1 were lexicalised by the exhaustive-control verb *shefa* ‘try’; those for Task 2 by the partial-control verb *xiangyao* ‘want’; and those for Task 4 by the non-control verb *shuo* ‘say’. These are typical verbs used in the literature to illustrate the respective (non-)control properties, making them ideal candidates for testing the hypothesised (non-)control-related displacement patterns.

Task 3 tested the generalisation that when the displaced phrase precedes a partial-control verb, the embedded subject must be unexpressed (Pattern C). Table 2 illustrates the four conditions, with the displacement phenomena and embedded-subject expression as the independent variables. The conditions are lexicalised using the partial-control verb *jueding* ‘decide’ (see Appendix A). All the conditions

<sup>15</sup>In total, the responses of 18 out of 106 participants were discarded. That means the responses of 88 participants were deemed valid responses for the subsequent data analysis.

	SUBJ unexpressed	SUBJ expressed
Focus fronting	SUBJ unexpressed + focus fronting (Condition A)	SUBJ expressed + focus fronting (Condition B)
Inner topicalisation	SUBJ unexpressed + inner topicalisation (Condition C)	SUBJ expressed + inner topicalisation (Condition D)

Table 2:  
2 × 2 factorial  
design of Task 3  
(Partial control –  
embedded SUBJ  
expression)

involved the configuration where the displaced phrase precedes the partial-control verb.

Task 5 tested the generalisation that for an object-control verb, the displaced phrase must not cross the object controller (Pattern E). Table 3 illustrates the four conditions, with the displacement phenomena and displacement positions as the independent variables. The conditions are lexicalised in Appendix A using the object-control verb *shuifu* ‘persuade’.

	Crossing OBJ controller	Not crossing OBJ controller
Focus fronting	Crossing OBJ controller + focus fronting (Condition A)	Not crossing OBJ controller + focus fronting (Condition B)
Inner topicalisation	Crossing OBJ controller + inner topicalisation (Condition C)	Not crossing OBJ controller + inner topicalisation (Condition D)

Table 3:  
2 × 2 factorial  
design of Task 5  
(OBJ controller)

#### Participants and apparatus

#### 4.6.2

The AJTs were designed as questionnaires using Qualtrics and distributed online to native Mandarin Chinese speakers. All 88 participants took part in Task 5, which was the only AJT found across Lists 1–8. Of the 88 participants, 48 of them also took part in Tasks 1 and 4, and 40 also participated in Tasks 2 and 3.<sup>16</sup> The participants were asked about their language background, for example, how old they

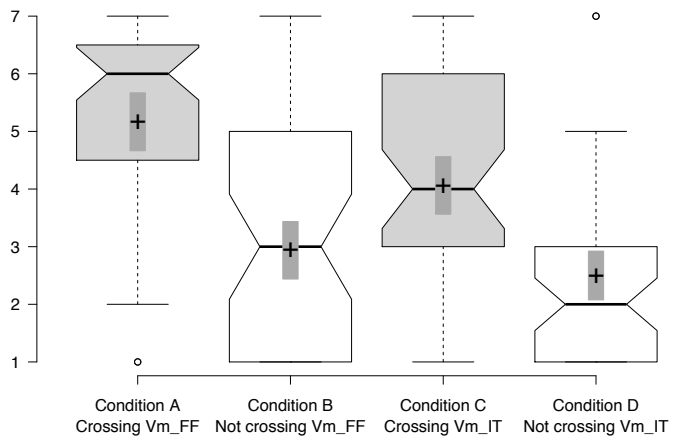
<sup>16</sup> We maintained a sample size of at least 37 participants per task to achieve 80% statistical power, following the calculation by Sprouse and Almeida (2012, p. 26) for medium-sized effect Likert-scale judgment tasks.

were when they started acquiring Mandarin Chinese, a self-report regarding their competence of the language, etc.<sup>17</sup> Participants were instructed to rate sentences on a 7-point Likert scale, accompanied by a plausible context. Clear instructions and examples were given before rating. A score of 1 indicated a completely unacceptable sentence, while a score of 7 indicated a perfectly natural sentence.

4.6.3 Results and preliminary trends

The results of the five AJTs are presented in Figures 1–5 in boxplots, created by the tool BoxPlotR (Spitzer *et al.* 2014). The notches represent the 95% confidence intervals of the medians. The black crosses indicate mean ratings. The grey areas around the crosses represent the 95% confidence intervals of the means. “FF” stands for focus fronting, and “IT” for inner topicalisation.

Figure 1:  
Results of  
Acceptability  
Judgment Task 1  
(Exhaustive  
Control)



Based on visual inspection, the overall trends supported Patterns A to E.<sup>18</sup> In addition, inner topicalisation tended to receive

<sup>17</sup> Participants who rated their language competence as “good” and started learning Mandarin Chinese before age six were included in the study. Some studies also administer competence tests to ensure native speaker status (e.g., Huang 2021), while others appear to rely on self-reported competence (e.g., Grano and Lasnik 2018; White and Grano 2014).

<sup>18</sup> As noted by one of the reviewers, the spread of data indicates speaker variation, which is common in any acceptability-judgment design, and it



*Control, inner topicalisation, and focus fronting*

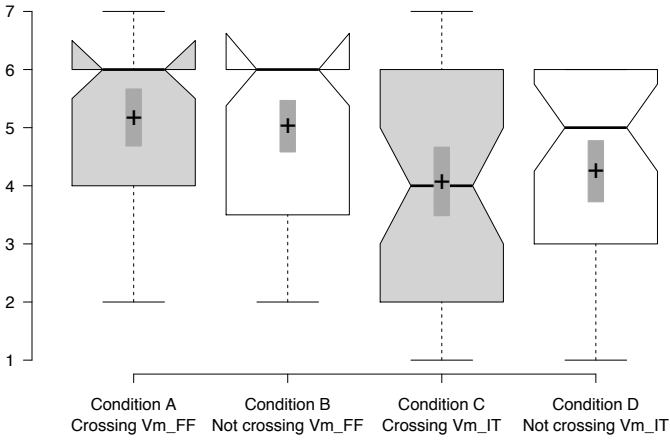


Figure 2:  
Results of  
Acceptability  
Judgment Task 2  
(Partial Control)

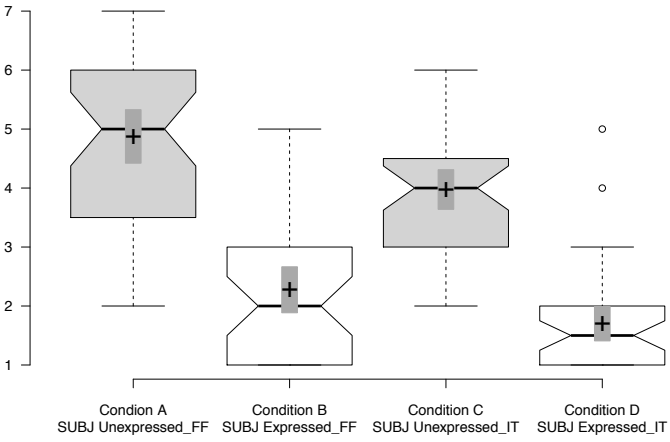


Figure 3:  
Results of  
Acceptability  
Judgment Task 3  
(Partial Control -  
embedded SUBJ  
expression)

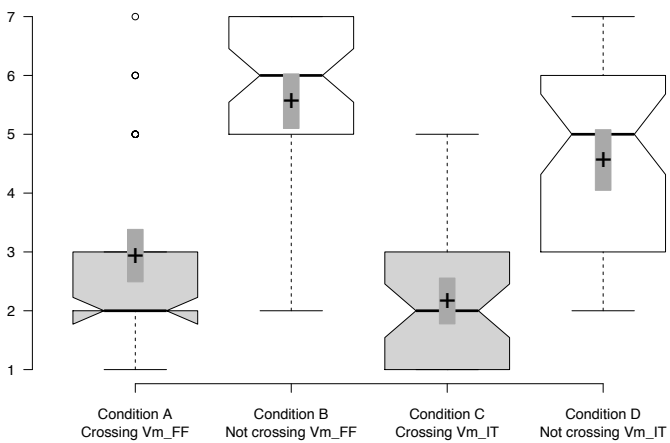
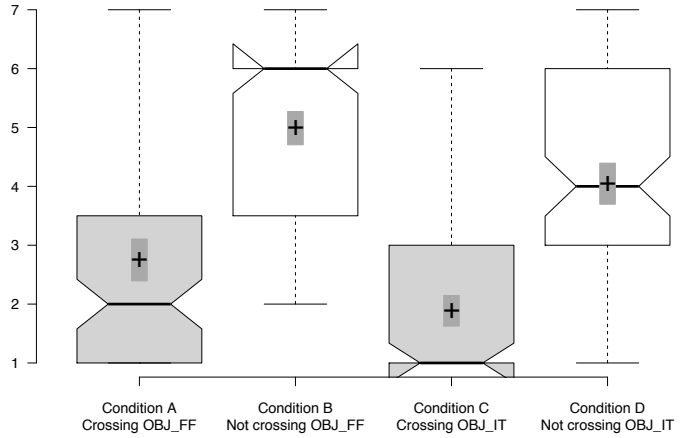


Figure 4:  
Results of  
Acceptability  
Judgment Task 4  
(Non-control)

Figure 5:  
Results of  
Acceptability  
Judgment Task 5  
(OBJ controller)



lower ratings than focus fronting in all AJTs. This observation has not been hitherto reported. Syntactic literature in general (e.g., Grano 2015; Huang 2018; Paul 2002, 2005) assumes both types to be equally acceptable by native speakers. We believe it is plausible for inner topicalisation to receive lower ratings than focus fronting in AJTs. Unlike focus fronting, inner topicalisation is not signalled by any overt markers, which means it could appear to participants as simply a construction that violates the usual SVO word order of Chinese. The fact that the AJTs were designed as written tasks could also be a reason for inner-topic constructions to be less favourably rated because inner topicalisation appears more often in the spoken form and less so in the written language, unlike focus fronting which is common in both spoken and written Chinese.<sup>19</sup> Despite these factors, it was still informative to compare experimental conditions of the same displacement phenomena.

is also common to accept that native speakers of the same language can have (slightly) different mental grammars. In what follows, we will employ mixed-effects analyses to identify which factors should be regarded as statistically significant and which should not. The statistical analyses support Patterns A to E, which are accounted for in the formal LFG analysis in Section 7.

<sup>19</sup>As suggested by one of the reviewers, to avoid this issue, future research on inner topicalisation may adopt a speech-based design via recordings.

We applied cumulative link mixed-effects models (ordinal regression) to analyse the results using the R package **ordinal** (Christensen 2020).<sup>20</sup> These models, which are also used in e.g., Huang (2021) and Bross (2019) for Likert-scale rating data, incorporated two main fixed effects: displacement positions and displacement phenomena for Tasks 1, 2, 4 and 5; and subject expression and displacement phenomena for Task 3. Random intercepts for participants and test items were included to account for random-variation effects.<sup>21</sup> An analysis of deviance, following Bross (2019), was conducted by fitting in each ordinal model using the R packages **RVAideMemoire** (Hervé 2022) and **car** (Fox and Weisberg 2019). The results, presented in Tables 4–13, are consistent with the predictions of the empirical generalisations (Patterns A–E). The results are consistent with the qualitative evidence examined in Sections 4.1–4.5. Future research may include a larger set of complementation verbs to be tested by AJTs using the same formats as the present study.

For Task 1, sentences with the displaced phrase remaining inside the embedded clause were rated significantly less acceptable than having the displaced phrase crossing the exhaustive-control predicate, in line with Pattern A. The analysis of deviance identified that displacement positions were a significant main effect. For Task 2, there was no significant difference in acceptability ratings between having the displaced phrase preceding vs following a partial-control predicate, although the former was rated slightly more acceptable. This result was in line with Pattern B. The analysis of deviance suggested that displacement positions were not a significant predictor of the ratings. For Task 3, constructions with an unexpressed embedded subject were significantly more acceptable than those with an expressed subject, in

---

<sup>20</sup> Following Bross (2019), we used z-transformed ratings to remove scale bias among participants. See Bross 2019, pp. 28–27 for a demonstration of how this step may help remove scale bias in a cumulative link mixed-effects model.

<sup>21</sup> Like Huang (2021), we tested and dismissed more complicated models that included random slopes and intercepts because they produced more random effects than data points, resulting in an insufficient number of observations to support the models.

line with Pattern C. An analysis of deviance indicated that embedded subject overttness was a significant main effect. For Task 4, the test sentences where the displaced phrase resides inside the complement clause were rated significantly more acceptable than those with the displaced phrase crossing the complementation verb, in line with Pattern D. An analysis of deviance revealed that displacement positions were a significant main effect. For Task 5, those constructions with the displaced phrase remaining inside the complement clause were rated to be significantly more acceptable than those with the displaced phrase crossing the object controller, in line with Pattern E. An analysis of deviance revealed that displacement positions were a statistically significant predictor.

Table 4:  
Mixed-effects  
regression  
analysis  
for Task 1  
(Exhaustive  
Control)

*Crossing V<sub>m</sub> and focus fronting as reference levels*

Condition	Estimate	Std error	z	p
<i>Displacement positions</i> Not crossing V <sub>m</sub>	-2.3920	0.3033	-7.886	3.11e-15 ***
<i>Displacement phenomena</i> Inner topicalisation	-0.9706	0.2670	-3.635	0.000278 ***

*Significance level: \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05*

Table 5:  
Analysis  
of deviance  
(Type II tests)  
for Task 1  
(Exhaustive  
Control)

	LR Chisq	Df	p
<i>Displacement positions</i>	22.1972	1	2.46e-06 ***
<i>Displacement phenomena</i>	7.7119	1	0.005486 ***

*Significance level: \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05*

Table 6:  
Mixed-effects  
regression  
analysis  
for Task 2  
(Partial Control)

*Crossing V<sub>m</sub> and focus fronting as reference levels*

Condition	Estimate	Std error	z	p
<i>Displacement positions</i> Not crossing V <sub>m</sub>	-0.009257	0.2791	-0.033	0.974
<i>Displacement phenomena</i> Inner topicalisation	-1.143342	0.2908	-3.931	8.46e-05 ***

*Significance level: \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05*

Control, inner topicalisation, and focus fronting

	LR Chisq	Df	p
<i>Displacement positions</i>	0.0014	1	0.969894
<i>Displacement phenomena</i>	9.4589	1	0.002101 **

Significance level: '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05

Table 7:  
Analysis  
of deviance  
(Type II tests)  
for Task 2  
(Partial Control)

*SUBJ expressed and focus fronting as reference levels*

Condition	Estimate	Std error	z	p
<i>SUBJ expression</i> <i>SUBJ unexpressed</i>	4.3567	0.4470	9.746	< 2e-16 ***
<i>Displacement phenomena</i> <i>Inner topicalisation</i>	-1.3669	0.2943	-4.645	3.41e-06 ***

Significance level: '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05

Table 8:  
Mixed-effects  
regression  
analysis for  
Task 3 (Partial  
Control – SUBJ  
expression)

	LR Chisq	Df	p
<i>SUBJ expression</i>	32.929	1	9.56e-09 ***
<i>Displacement phenomena</i>	10.497	1	0.001196 **

Significance level: '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05

Table 9:  
Analysis  
of deviance  
(Type II tests)  
for Task 3  
(Partial Control –  
SUBJ expression)

*Crossing V<sub>m</sub> and focus fronting as reference levels*

Condition	Estimate	Std error	z	p
<i>Displacement positions</i> <i>Not crossing V<sub>m</sub></i>	3.3942	0.3572	9.503	< 2e-16 ***
<i>Displacement phenomena</i> <i>Inner topicalisation</i>	-1.2781	0.2680	-4.768	1.86e-06 ***

Significance level: '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05

Table 10:  
Mixed-effects  
regression  
analysis  
for Task 4  
(Non-control)

	LR Chisq	Df	p
<i>Displacement positions</i>	43.718	1	3.793e-11 ***
<i>Displacement phenomena</i>	12.961	1	0.000318 ***

Significance level: '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05

Table 11:  
Analysis  
of deviance  
(Type II tests)  
for Task 4  
(Non-control)

Table 12:  
Mixed-effects  
regression  
analysis  
for Task 5  
(OBJ controller)

Condition	Estimate	Std error	z	p
<i>Displacement positions</i> Not crossing OBJ controller	2.7767	0.3604	7.704	1.32e-14 ***
<i>Displacement phenomena</i> Inner topicalisation	-1.1447	0.2508	-4.564	5.03e-06 ***

Significance level: \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05

Table 13:  
Analysis  
of deviance  
(Type II tests)  
for Task 5  
(OBJ controller)

	LR Chisq	Df	p
<i>Displacement positions</i>	42.627	1	6.624e-11 ***
<i>Displacement phenomena</i>	14.491	1	0.0001408 ***

Significance level: \*\*\*\* 0.001 \*\*\* 0.01 \*\* 0.05

## 5

PRE-THEORETICAL INSIGHTS  
OF ICH SIGNATURE AND A LEXICALIST  
APPROACH TO BRIDGE VERBS

Wurmbrand and Lohninger (2019) identify three types of complementation that are cross-linguistically available, namely, Proposition (*claim-type*), Situation (*decide-type*) and Event (*try-type*). Proposition complements typically involve speech and epistemic contexts; Situation complements are typically related to emotive and irrealis contexts; and Event complements typically involve implicative and strong attempt contexts. These three types of complementation form the Implicational Complementation Hierarchy (ICH). The Proposition complement is ranked as the most independent/least transparent among the three, whereas the Event complement is regarded as the least independent/most transparent. According to Wurmbrand and Lohninger (2019, pp. 5–6), “independence” is manifested by properties such as the interpretation and overtness of an embedded subject, while “transparency” is signalled by the permeability for dependency relations.

ICH relates to control relations: Event-complements often involve exhaustive-control verbs (e.g., *try*, *manage*); Situation-complements

often involve partial-control verbs (e.g., *decide, want*); and Proposition-complements often involve non-control verbs (e.g., *claim, say*). Applying ICH's characteristic of "independence" to control relations, exhaustive control exhibits the lowest degree of independence by enforcing strict coreference between the controller and controllee. Also, cross-linguistically, exhaustive-control verbs often require the embedded subject to be unexpressed (see, e.g., Stiebels 2007). Non-control relation signals the highest degree of independence by allowing the embedded subject to be freely interpretable. Partial control occupies the middle ground, with the controller and controllee forming a subset relation. Applying ICH's notion of "transparency" to inner topicalisation and focus fronting, obligatory displacement of inner topic/focused phrase into the matrix clause manifests the highest degree of permeability of displacement-dependency relation across the clausal boundary, while obligatory retainment of inner topic/focused phrase in the complement clause signals the lowest degree of permeability.

Wurmbrand and Lohninger (2019) discusses the ICH Signature, which governs the distribution of a property across the three complementation types. According to the ICH Signature, when a property (P) distinguishes among the three types of complements, the Proposition complement and Event complement illustrate opposite values, whereas the Situation complement either allows both values or sides with one of them. By examining a range of cross-linguistic patterns pertinent to complementation (e.g., finiteness, clitic climbing, complementiser distribution), Wurmbrand and Lohninger (2019) conclude that there are important universal hierarchical effects: in a given language, if the Situation complement possesses a transparency property, the Event complement will also possess it; if the Proposition complement possesses a transparency property, both Situation complement and Event complement will also possess it. Placing inner topicalisation and focus fronting in the wider picture of ICH, our observed empirical patterns (Patterns A, B and D) align with the predictions of the ICH Signature. Focusing on subject control, Table 14 illustrates the alignment patterns, with "P" standing for a transparency property.

Patterns A to E essentially suggest that inner topicalisation and focus fronting correlate with complement control – a lexically determined phenomenon from the perspective of LFG (Bresnan 1982; Bres-

Table 14:  
ICH Signature  
(Wurmbrand and  
Lohninger 2019),  
control relations,  
and Chinese  
inner  
topicalisation /  
focus fronting

	Proposition ( <i>claim-type</i> )	Situation ( <i>decide-type</i> )	Event ( <i>try-type</i> )
	most independent $\longleftrightarrow$ least independent least transparent $\longleftrightarrow$ most transparent		
ICH Signature	-P	$\pm$ P	+P
Control relation	Non-control	Partial control	Exhaustive control
Inner Top. / Focus Front.	Not crossing $V_m$	Crossing $V_m$ or not crossing $V_m$	Crossing $V_m$

nan *et al.* 2016; Dalrymple *et al.* 2019). Another displacement phenomenon – the “bridge-verb effect” (Erteschik 1973) – is also known to be lexically determined in the LFG literature. This phenomenon sheds light on the issues at hand. In English, bridge verbs (e.g., *say*, *think*, *report*, *announce*) are said to allow extraction out of their clausal complement in contrast to non-bridge verbs (e.g., *whisper*, *stammer*, *dictate*, *snarl*), whose complement clause forms an island impermeable to extraction, as exemplified in (31):

- (31) a. Thomas, we said/thought that Sarah saw.  
 b. \*Thomas, we whispered/stammered that Sarah saw.

The bridge-verb effect has been analysed in different ways: some approach it from the perspective of information structure (e.g., Ambridge and Goldberg 2008), and some from the perspective of verb frequency (e.g., Liu *et al.* 2022), etc.<sup>22</sup> In LFG, the bridge-verb effect has been analysed syntactically using a lexicalist mechanism (Butt *et al.* 1999; Dalrymple *et al.* 2019). Dalrymple *et al.* (2019, pp. 226–227) propose that the distinction between bridge and non-bridge verbs should not be reflected in the grammatical function of their complement clause; instead, some additional feature is lexically imposed by the non-bridge verb on its functional structure. The feature interacts with a mathematically well-defined extraction formula encoded on a phrase-structural rule to render its complement clause an unextractable island. While more will be said about the LFG formalism,

<sup>22</sup>See Huang *et al.* (2022) for an experimental evaluation of some of these non-syntactic approaches.



what is important to note here is that LFG's approach to the bridge-verb effect is a lexicalist proposal which does not hypothesise any clause-size differences in the complement clause; rather, the effect is directly regulated by the verb. This captures the basic insight that the differences between (31a) and (31b) lie in the differences of the complementation verbs, rather than the size of their complement clause.

LFG's approach to the bridge-verb effect has offered insights into how we can model the interaction between complement control and inner topicalisation/focus fronting. Section 7 will demonstrate how LFG's bridge-verb mechanism can be incorporated into the modelling of inner topicalisation/focus fronting, enabling the complementation verb to regulate displacement patterns directly without positing any clause-size distinctions in the complement clause, contra restructuring proposals. Before then, note that we will deal with a tripartite distinction of extraction patterns (rather than a bipartite one): (i) the displaced phrase crossing the matrix predicate; (ii) the displaced phrase remaining in the complement clause; (iii) the displaced phrase either crossing the matrix predicate or remaining in the complement clause. Simply assigning a bridging feature cannot capture the tripartite distinction, so some additional formal mechanism will be needed.

The next section will briefly introduce the LFG formalism as well as how LFG handles control. Section 7 will devise a formal LFG mechanism to model inner topicalisation and focus fronting.

## LFG: FORMALISM, CONTROL, AND BRIDGE-VERB MECHANISM

6

LFG is a constraint-based formal grammatical theory, first developed by Joan Bresnan and Ronald Kaplan in the 1970s (Kaplan and Bresnan 1982). This formalism is presented in detail by e.g., Bresnan *et al.* 2016 and Dalrymple *et al.* 2019. Of crucial importance is the idea of a parallel architecture, where different types of linguistic information are represented as distinct formal structures with their own notations, interrelated by projection functions.

This paper focuses on two formal structures: the constituent structure (c-structure) and functional structure (f-structure), which are syntactic structures. The c-structure takes the form of a labelled tree to encode constituency, dominance, and linear order. A c-structure is formed by phrase-structure rules, which loosely observe a version of the X'-theory (Jackendoff 1977; see also Bresnan *et al.* 2016, pp. 101–111).<sup>23</sup> The f-structure takes the form of an attribute-value matrix, encoding grammatical functions (e.g., subject SUBJ, object OBJ, adjunct ADJ) and features (e.g., person, number, tense, aspect). The c- and f-structures are related by a projection function, mapping c-structural nodes to their corresponding f-structures. The f-structure is built up using the functional information encoded in annotated c-structural rules and lexical entries. See, e.g., Bresnan *et al.* (2016, pp. 54–58) for the solution algorithm for building up the f-structure, which we will skip here. The f-structure is the locus of explanation for control relations.

Since Bresnan 1982, LFG has assumed two main model-theoretic control mechanisms, namely functional control and anaphoric control (see also Andrews 1982; Bresnan *et al.* 2016, pp. 286–323; Dalrymple *et al.* 2019, pp. 545–601; Mohanan 1983). Functional control involves f-structural identity such that the controller and controllee share the same f-structure; on the other hand, in anaphoric control, the controllee is represented as a pronominal entity which is syntactically independent of the controller. A number of LFG studies represent exhaustive control as functional control (e.g., Asudeh 2005; Bresnan 1982; Bresnan *et al.* 2016), where the complete identity between the controller and controllee is attributed to a structure-sharing mechanism. We will follow this approach in this paper (see Section 7).<sup>24</sup> Regarding partial control, past research differs on whether partial control should be represented as functional control (Asudeh 2005) or a subtype of anaphoric control known as “quasi-obligatory anaphoric

---

<sup>23</sup> See also Lowe and Lovstrand (2020) for an alternative LFG phrase-structure theory that incorporates insights from Bare Phrase Structure. This paper will stick to the version of X'-theory commonly found in LFG studies.

<sup>24</sup> An alternative view is that exhaustive control involves obligatory anaphoric control (Dalrymple *et al.* 2019, pp. 545–601), where the enforced identity between the controller and controllee is attributed to a semantic constraint.

control” (Haug 2013, 2014). Both proposals involve some semantic constraints for modelling partial coreference. Asudeh (2005, p. 504) incorporates a subsumption operator in the predicate-logic side of a partial-control verb’s meaning constructor to capture the relation between the controller and controllee, specifying that the controller is either semantically the same as or part of the controllee. Haug (2013) posits a semantic locality constraint, capturing the nature of control as a logophoric-binding relation between the controller (logocentre) and controllee (logophor). Haug’s (2013) proposal has been adopted by Dalrymple *et al.* (2019).<sup>25</sup> In this paper, we will adopt an anaphoric-control approach to model partial control, aligning us more with Haug (2013). However, because this paper focuses on syntactic structures, we will skip semantic constraints in the analysis. As will be shown in Section 7.3, our anaphoric approach to partial control includes two attribute-value pairs, <P\_CONTROL, CONTROLLER> and <P\_CONTROL, CONTROLLEE>, in the f-structure to clearly indicate which grammatical function serves as the controller and which serves as the controllee. Note that while these attribute-value pairs are useful indicators of partial control, the actual modelling of the partial conference (where the entity denoted by the controller is a subset of the entities denoted by the controllee) takes place in the semantics as discussed by Haug (2013), from which we have abstracted away due to the syntactic focus of this paper.<sup>26</sup>

---

<sup>25</sup> Since Dalrymple *et al.* (2019) also treat exhaustive control as a type of anaphoric control, they regard both exhaustive control and partial control as anaphoric in the f-structure but differ significantly in the formal semantics to capture the different empirical properties embodied by these two control types. In other words, in LFG, it is theoretically possible to treat both control types uniformly in the syntax and model their differences in the semantics. That being said, Lam (2023) draws on in-depth empirical data and argues that, even within the exhaustive-control class in Chinese, not all of the verbs can be said to behave the same syntactically; while some involve functional control, others are best analysed as involving anaphoric control. This paper will not go into such details.

<sup>26</sup> In Section 7, the P\_CONTROL attribute will be useful in stating implicational constraints for partial control scenarios when we devise a template for all Chinese complementation verbs.

As was mentioned previously, our proposal assimilates LFG’s bridge-verb mechanism. In Dalrymple *et al.* 2019, pp. 226–228, non-bridge verbs specify that their complement clause contains the attribute-value pair <LDD, – > (where LDD stands for “long-distance dependency”). The extraction formula for long-distance dependency, which is encoded on a phrase-structure rule, imposes a condition on the extraction path such that the path must not contain <LDD, – >. Example (33) contains the lexical entry of the non-bridge verb *stammer*, an ill-formed sentence of *stammer*, and its invalid f-structure. DIS is the displacement function and its value is a set, whose member is related to the embedded OBJ inside the complement-clause function COMP (both notated by the same boxed number); as such, the f-structure models the topicalisation of the embedded OBJ *Thomas* to the matrix-clause level.<sup>27</sup> Example (32) shows the extraction path (*f* COMP OBJ) for the topicalisation of *Thomas* in (33), where *f* is the outermost f-structure, relating the topicalised phrase to the embedded OBJ function.

(32) The extraction path for (33) with an off-path constraint is

$$(f \text{ COMP OBJ})_{(\rightarrow \text{LDD}) \neq -}$$

(33) \**Thomas, we stammered that Sarah saw.*

The lexical entry of *stammer* is:

$$\textit{stammer} \quad \text{V} \quad (\uparrow \text{PRED}) = \text{'STAMMER <SUBJ, COMP>'}$$

$$(\uparrow \text{COMP LDD}) = -$$

Its invalid f-structure is:

$$f \left[ \begin{array}{l} \text{DIS} \quad \left\{ \boxed{1} \left[ \text{PRED 'THOMAS'} \right] \right\} \\ \text{PRED} \quad \text{'STAMMER <SUBJ, COMP>'} \\ \text{SUBJ} \quad \left[ \text{PRED 'PRO'} \right] \\ \text{COMP} \quad \left[ \begin{array}{l} \text{PRED 'SEE <SUBJ, OBJ>'} \\ \text{SUBJ} \quad \left[ \text{PRED 'SARAH'} \right] \\ \text{OBJ} \quad \boxed{1} \\ \text{LDD} \quad - \end{array} \right] \end{array} \right]$$

<sup>27</sup>There are two clausal functions in LFG. COMP is a closed clausal function used in anaphoric control. XCOMP is an open clausal function associated with functional control (Section 7.2).

Encoded beneath the extraction path is the negative off-path constraint ( $\rightarrow$  LDD)  $\neq -$ , whose right arrow stands for the value of the attribute COMP.<sup>28</sup> The off-path constraint forbids COMP from containing  $\langle$ LDD,  $-$  $\rangle$ . The f-structure in (33) cannot satisfy this off-path constraint since its COMP contains  $\langle$ LDD,  $-$  $\rangle$ , specified by the lexical entry of *stammer* in (33).

In the next section, we will see how the bridge-verb mechanism can be incorporated to model the interaction among control, inner topicalisation, and focus fronting.

## LFG FORMAL ANALYSIS OF INNER TOPICALISATION AND FOCUS FRONTING

7

This section will provide a formal LFG analysis of inner topicalisation and focus fronting, capturing their interaction with control and complementation. It is a non-movement and lexicalist analysis, placing emphasis on the role of the lexicon in governing the patterns. This is in contrast to past restructuring proposals, which rely on clause-sized differences that are not supported by independent syntactic evidence. The analysis assimilates LFG's bridge-verb mechanism (Section 6) and involves the lexicon introducing the feature PS\_LDD (acronym for "Post-Subject (position) Long-Distance Dependency"), which is reminiscent of Dalrymple *et al.*'s (2019) LDD bridging feature. The interaction between the PS\_LDD bridging feature and an annotated phrase-structural rule provides the formal means for the embedded object to appear in the matrix clause while keeping the clausal boundary intact in both c- and f-structures. Additional formal devices will be employed to obtain the tripartite distinction discussed in Section 5.<sup>29</sup>

---

<sup>28</sup> See Dalrymple *et al.* e.g., 2019, pp. 225–230 and Börjars *et al.* 2019, p. 145 for more information on how to use off-path constraints.

<sup>29</sup> Our constraint-based model characterises a binary distinction between "grammatical" and "ungrammatical" structures, similar to many theoretical linguistic analyses, rather than a gradient distinction that may be more closely matched with the gradient ratings gathered from the native speakers in the acceptability-judgment tasks. In fact, it remains a debatable issue in the field of

7.1 *Phrase-structural rules with functional annotations*

Our formal grammar contains, among others, several phrase-structural rule sets listed in (34) to (37) that are particularly relevant to modelling inner topicalisation and focus fronting. These rules are annotated with functional constraints.<sup>30</sup>

(34) **IP and I' rules**

$$\begin{aligned} \text{IP} &\rightarrow \left( \begin{array}{c} \text{DP} \\ (\uparrow \text{SUBJ}) = \downarrow \end{array} \right) \begin{array}{c} \text{I}' \\ \uparrow = \downarrow \end{array} \\ \text{I}' &\rightarrow \left\{ \begin{array}{c} \text{DP} \\ \downarrow \in (\uparrow \text{DIS}) \\ \text{PS\_LDD-PATH} \end{array} \begin{array}{c} \text{I}' \\ \uparrow = \downarrow \end{array} \mid \left( \begin{array}{c} \text{I} \\ \uparrow = \downarrow \end{array} \right) \begin{array}{c} \text{VP} \\ \uparrow = \downarrow \end{array} \right\} \\ \text{PS\_LDD-PATH} &\equiv (\uparrow (\{ \text{XCOMP} \mid \text{COMP} \} \{ \text{XCOMP} \mid \text{COMP} \}^* ) \text{OBJ}) = \downarrow \\ &\quad (\rightarrow \text{PS\_LDD}) = \text{c} + \end{aligned}$$

(35) **Complex-category IP<sub>[-PS\_LDD]</sub> and I'<sub>[-PS\_LDD]</sub> rules**

$$\begin{aligned} \text{IP}_{[-\text{PS\_LDD}]} &\rightarrow \begin{array}{c} \text{I}'_{[-\text{PS\_LDD}]} \\ \uparrow = \downarrow \end{array} \\ \text{I}'_{[-\text{PS\_LDD}]} &\rightarrow \left( \begin{array}{c} \text{I} \\ \uparrow = \downarrow \end{array} \right) \begin{array}{c} \text{VP} \\ \uparrow = \downarrow \end{array} \end{aligned}$$

experimental syntax whether recognising acceptability judgment as a gradient factor in empirical experiments entails accepting grammaticality as a gradient notion in formal language modelling. See, e.g., Goodall 2021a. From the perspective of Bader and Häussler (2010, p. 276), while it is one thing to accept the gradience of acceptability judgments, it is another thing to accept the notion of gradient grammaticality. That being said, in our acceptability-judgment tasks, we employed experimental paradigms that enabled us to measure whether a potential governing factor is statistically significant or not based on p-values. Such decisions of statistical significance are also binary in nature. Another oft-mentioned issue in experimental syntax is the difference between “grammaticality” and “acceptability”. A discussion about this issue can be found in Goodall 2021b.

<sup>30</sup>We follow the approach in Dalrymple *et al.* 2019, where the constituents on the right-hand side of a phrase-structural rule are not by default optional and any optionality of constituents is marked by parentheses (...). Curly brackets indicate a disjunction of phrase-structure categories with the possibilities separated by a vertical bar {...|...}.

(36) VP and V' rules<sup>31</sup>

$$\begin{array}{l}
 \text{VP} \rightarrow \begin{array}{c} \text{V}' \\ \uparrow=\downarrow \end{array} \\
 \text{V}' \rightarrow \left\{ \left\{ \begin{array}{c} \text{PRT} \\ \uparrow=\downarrow \end{array} \mid \begin{array}{c} \text{AdvP} \\ \downarrow \in (\uparrow \text{ADJ}) \end{array} \right\} \text{V}' \right. \\
 \left. \mid \begin{array}{c} \text{V} \\ \uparrow=\downarrow \end{array} \left( \begin{array}{c} \text{DP} \\ (\uparrow \text{OBJ})=\downarrow \end{array} \right) \left( \left\{ \begin{array}{c} \text{IP} \\ (\uparrow \{\text{XCOMP} \mid \text{COMP}\})=\downarrow \end{array} \mid \begin{array}{c} \text{IP}_{[-\text{PS\_LDD}]} \\ (\uparrow \text{XCOMP})=\downarrow \end{array} \right\} \right) \right\}
 \end{array}$$

## (37) DP-adjoining rule

$$\text{DP} \rightarrow \begin{array}{c} \text{AdvP} \quad \text{DP} \\ (\uparrow \text{SPEC})=\downarrow \quad \uparrow=\downarrow \end{array}$$

Rule set (34) contains an I'-adjoining rule licensing the structural position where an inner topic or focused phrase (bearing the DP category) surfaces.<sup>32</sup> Chinese SUBJ in general occupies a pre-verbal position (see e.g., Li and Thompson 1989). With SUBJ being associated with the Spec-IP position (see e.g., Che and Bodomo 2018; Her 2009), an inner topic or focused phrase (lower than matrix subject but above the matrix predicate) is adjoined to I'. Although external topicalisation is not the issue here, further evidence that a Chinese inner topic or focused phrase occupies a position within the IP domain (rather than the CP domain) can be adduced from the structural position of external topicalisation inside a complement clause. According to Bresnan *et al.* (2016, pp. 16–17) and Dalrymple *et al.* (2019, p. 659), an English (external) topic inside the complement clause is adjoined to IP as is derived from the pattern in (38a), where the topic *Chris* appears after the complementiser *that* and before the embedded subject *we* rather than preceding the complementiser; thus, motivating an IP-adjoining position rather than the Spec-CP position.

<sup>31</sup>The V' rule contains both disjunctive and optional phrase-structure categories. As such, V' is capable of branching into one of the following: (i) PRT V'; (ii) AdvP V'; (iii) V; (iv) V DP; (v) V IP; (vi) V IP<sub>[-PS\_LDD]</sub>; (vii) V DP IP; (viii) V DP IP<sub>[-PS\_LDD]</sub>.

<sup>32</sup>In this paper, we assume that Chinese nominal phrases are DPs rather than NPs. See Börjars *et al.* 2018 and Her 2012 for further discussion on the internal structure of Chinese nominal phrases from LFG perspectives.

- (38) a. Matty thinks that [Chris] we like.  
 (Dalrymple *et al.* 2019, p. 659)
- b. xiaoming renwei shuo [zhe-ben shu] ta hui xihuan  
 Xiaoming think COMP this-CL book 3SG will like  
 ‘Xiaoming thinks that he will like this book.’
- c. xiaoming renwei shuo ta [zhe-ben shu] hui xihuan  
 Xiaoming think COMP 3SG this-CL book will like  
 ‘Xiaoming thinks that he will like this book.’

Likewise, as shown in (38b), a Chinese external topic inside the complement clause appears after the complementiser *shuo* and before the embedded subject *ta* ‘he’ rather than preceding the complementiser. Thus, the Chinese external topic should also be placed in an IP-adjoining position rather than the Spec-CP position. As the external-topic position is associated with the IP domain, this in turn suggests that a Chinese inner topic (or focused phrase) should not be analysed as belonging to the higher CP domain.<sup>33</sup> Assuming that a modal auxiliary occupies the I position, given that the inner topic in (38c) precedes the future modal *hui* ‘will’, it must occur in the IP domain (above I) rather than the lower VP domain.<sup>34</sup> Therefore, in our treatment, Chinese external topic, subject, inner topic and focused phrase are all constituents of the IP domain.

Encoded below DP of the I'-adjoining rule in (34) are two lines of functional annotation. The first line states that the f-structure corresponding to DP maps onto a member of the DIS set in the f-structure. DIS is adopted from Dalrymple *et al.* 2019, p. 37 as a function of long-distance dependency borne by a fronted phrase. Any member of DIS must be integrated into an f-structure built up around a predicate via f-structural sharing (or anaphoric binding), establishing a dependency relationship between a member of DIS and a within-clause function. This formal setup is governed by a well-formedness principle – the

<sup>33</sup> Our approach differs from Paul's (2002; 2005) regarding the functional projections for hosting topic and focused phrases. Working in a different analytic framework, Paul's (2002; 2005) phrase-structural treatment is different from the LFG phrase-structure theory adopted in our paper.

<sup>34</sup> We differ from Ernst and Wang's (1995) proposal where an inner topic is adjoined to VP.



“Extended Coherence condition” (Zaenen 1980; see also Bresnan *et al.* 2016, pp. 62–63; Dalrymple *et al.* 2019, p. 653). The second line contains an extraction formula PS\_LDD-PATH, which presides over a set of possible paths through the f-structure to the within-clause function (OBJ) of the displaced phrase. The asterisk \* in the path is a Kleene star operator, indicating that there can be zero to infinite instances of XCOMP or COMP. Here, functional uncertainty is invoked to capture the different possibilities. The formal definition of functional uncertainty is cited from Kaplan and Zaenen 1989, p. 147:

(39) Functional uncertainty

Suppose  $\alpha$  is a (possibly infinite) set of strings.  $(f \alpha) = \nu$  holds if and only if  $((f s) \text{Suff}(s, \alpha)) = \nu$  for some symbol  $s$ , where  $\text{Suff}(s, \alpha)$  is a set of suffix strings  $y$  such that  $sy \in \alpha$ .

Applying this definition to PS\_LDD-PATH, a possible extraction path is one of the potentially infinite elements in the set {OBJ, XCOMP OBJ, COMP OBJ, XCOMP COMP OBJ, COMP XCOMP OBJ, XCOMP COMP XCOMP OBJ...}, where each of the possible paths must end with OBJ – the within-clause function borne by the displaced phrase. Note that if the path starts with a clausal function (either XCOMP or COMP), the f-structure of this function must contain the attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$ , which is the bridging attribute-value pair for licensing the extraction of an inner topic or focused phrase into the matrix clause. This requirement is imposed via an off-path constraint  $(\rightarrow \text{PS\_LDD}) =_c +$  on the beginning clausal function of the extraction formula PS\_LDD-PATH but does not apply to any subsequent clausal functions. From Section 7.2 onwards, we will see how the extraction formula works with language examples.

Note that there is a competing version of PS\_LDD-PATH as shown in (40), where each of the clausal functions (if any) has to satisfy the off-path equation  $(\rightarrow \text{PS\_LDD}) =_c +$ . Based on the data from Section 7.2 to Section 7.5, it is not possible to reject this competing version. However, when we proceed to complex-level embedding in Section 7.7, there is evidence to adjudicate that the extraction formula in (34) is the correct one.

(40) A competing (but incorrect) version of PS\_LDD-PATH

$$\text{PS\_LDD-PATH} \equiv (\uparrow \{ \text{XCOMP} | \text{COMP} \}^* \text{OBJ}) = \downarrow_{(\rightarrow \text{PS\_LDD}) =_c +}$$

Rule set (35) contains the complex category  $IP_{[-PS\_LDD]}$  and a set of its associated c-structural rules.<sup>35</sup> There is no  $I'$ -adjoining rule available inside the set of  $IP_{[-PS\_LDD]}$ -associated rules. Because an  $I'$ -adjoining rule is essential for licencing the structural position of an inner topic or focused phrase, the absence of this rule would render the formal grammar incapable of parsing a sentence where the inner topic or focused phrase appears inside the  $IP_{[-PS\_LDD]}$  domain; thus, such a sentence is flagged as ungrammatical. As will be discussed in Section 7.2, the  $IP_{[-PS\_LDD]}$ -associated rules are essential for the displacement patterns of exhaustive subject control verbs.

In focus fronting, because a focused phrase is introduced by a focus marker such as *lian* ‘even’, there needs to be an additional AdvP node for the marker, whose structural position is licensed by the DP-adjoining rule in (37). Given the functional annotation  $(\uparrow \text{SPEC}) = \downarrow$  on the AdvP node, the f-structure associated with its parent’s node DP contains the feature SPEC. In LFG, a SPEC feature is reserved for elements in a nominal phrase which carry “specifying” properties rather than serving modifying purposes (Dalrymple *et al.* 2019, pp. 83–84). The focus marker *lian* ‘even’ serves the purpose of specifying that a phrase is a focused phrase in addition to any modifying meaning it may add.

## 7.2

*Exhaustive subject control (Pattern A)*

As an illustration, (41) displays the lexical entry of the exhaustive-control verb *shefa* ‘try’, instantiating functional control.  $(\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ})$  is the functional-control equation, stating that the matrix subject and the subject in the complement clause XCOMP share the same f-structure. The lexical entry contains some crucial constraints responsible for the verb’s behaviour in inner topicalisation and focus fronting. *Shefa* ‘try’ assigns the bridging attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$  such that it is possible to extract the displaced phrase into the matrix clause. It is important to prevent the displaced phrase from residing in the complement clause.  $\text{CAT}((\uparrow \text{XCOMP}), \{IP_{[-PS\_LDD]}\})$  achieves this. To understand this constraint, note that, in LFG, subcategorisation requirements are stated

<sup>35</sup> Complex categories are detailed in e.g., Dalrymple *et al.* 2019, p. 250.

in f-structural terms (e.g., a verb subcategorising for SUBJ, OBJ, etc.). That being said, it is possible to impose c-structural categorical requirements on the f-structure of a grammatical function. For example, *shefa* ‘try’ subcategorises for XCOMP as its complement clause. The constraint  $CAT((\uparrow \text{XCOMP}), \{\text{IP}_{[-\text{PS\_LDD}]}\})$  uses the CAT predicate to impose a categorical requirement on the f-structure of this XCOMP such that the category of one of the nodes is constrained to be  $\text{IP}_{[-\text{PS\_LDD}]}$ . The formal definition of the CAT predicate is cited from Dalrymple *et al.* 2019, p. 250 (see also Crouch *et al.* 2011), using LFG’s projection architecture:

(41) Lexical entry of *shefa* ‘try’:

*shefa* ‘try’    V     $(\uparrow \text{PRED}) = \text{‘TRY <SUBJ, XCOMP >’}$   
 $(\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ})$   
 $(\uparrow \text{XCOMP PS\_LDD}) = +$   
 $CAT((\uparrow \text{XCOMP}), \{\text{IP}_{[-\text{PS\_LDD}]}\})$

(42) CAT predicate

$CAT(f, C)$  if and only if  $\exists n \in \phi^{-1}(f): \lambda(n) \in C$   
 $CAT(f, C)$  is true if and only if there is some node  $n$  that corresponds to  $f$  via the inverse  $\phi$  correspondence ( $\phi^{-1}$ ) whose label ( $\lambda$ ) is in the set of categories.

The set of  $\text{IP}_{[-\text{PS\_LDD}]}$  rules with the CAT predicate means that *shefa* ‘try’ is forced to select for a complement clause of the  $\text{IP}_{[-\text{PS\_LDD}]}$  category, whose domain cannot host an inner topic or focused phrase. In other words, the only structural position for its inner topic or focused phrase is the I’-adjoining position in the matrix clause.

Sentence (43) is an example of *shefa* ‘try’.

(43) xiaoming<sub>i</sub> [zhe-xiang gongzuo] shefa  $\emptyset_{i/*j}$  (jinkuai)  
 Xiaoming this-CL task try  $\emptyset$  soon  
 wancheng  
 finish  
 ‘Xiaoming tries to finish this task soon.’

Parsing this sentence will result in the c-structure in Figure 6 and f-structure (44), where for simplicity we have omitted any adjuncts.



$$(44) \left[ \begin{array}{l} \text{PREL} \\ \text{DIS} \\ \text{f} \\ \text{SUBJ} \\ \text{XCOMP} \end{array} \left[ \begin{array}{l} \text{PREL} \quad \text{'TRY <SUBJ, XCOMP>'} \\ \left\{ \begin{array}{l} \text{PREL} \quad \text{'TASK'} \\ \text{DEF} \quad + \\ \text{DEIXIS} \quad \text{PROXIMAL} \end{array} \right\} \\ \text{SUBJ} \quad \text{②} \left[ \text{PREL} \quad \text{'XIAOMING'} \right] \\ \left[ \begin{array}{l} \text{PREL} \quad \text{'FINISH <SUBJ, OBJ>'} \\ \text{SUBJ} \quad \text{②} \\ \text{OBJ} \quad \text{①} \\ \text{PS\_LDD} \quad + \end{array} \right] \end{array} \right. \right]$$

In the c-structure (Figure 6), we display the functional information contributed by the lexicon under the leaves of the c-structural tree.<sup>36</sup>

In the c-structure, the inner topic *zhe-xiang gongzuo* ‘this task’ is adjoined to *I'*. This structural position is licensed by the *I'*-adjoining rule in (34). In the f-structure, the inner topic is a member of the DIS set, which is a function at the matrix-clause level, and its extraction path is (*f* XCOMP OBJ), where *f* is the f-structure of the matrix clause. There is a dependency relationship between a member of the DIS set and the within-clause function OBJ in the form of f-structural sharing, which is licensed by the long-distance dependency equation PS\_LDD-PATH notated on the *I'*-adjoining rule in (34). XCOMP contains the bridging attribute-value pair <PS\_LDD, +>, satisfying the off-path equation ( $\rightarrow$  PS\_LDD) = <sub>c</sub> + in PS\_LDD-PATH. This attribute-value pair is specified by the lexical entry of *shefa* ‘try’ in (41) via the defining equation ( $\uparrow$  XCOMP PS\_LDD) = +. The f-structure shows structural sharing between the matrix SUBJ and embedded SUBJ due to functional control.

- (45) \*xiaoming<sub>i</sub> shefa  $\emptyset_{i/*j}$  [zhe-xiang gongzuo] wancheng  
 Xiaoming try  $\emptyset$  this-CL task finish  
 ‘Xiaoming tries to finish this task.’

On the other hand, (45) is flagged by the formal grammar as an ill-formed construction, for which no solution can be produced due

<sup>36</sup> From Section 7.3 onwards, we will skip the display of the lexical information in c-structures to reduce notational clutter.

to conflicts of constraints arising from a series of calculations as follows. In (45), the inner topic appears inside the complement clause XCOMP. The lexical entry of *shefa* ‘try’ in (43) contains the constraint  $CAT((\uparrow XCOMP), \{IP_{[-PS\_LDD]}\})$ , which forces XCOMP to be associated with  $IP_{[-PS\_LDD]}$ .<sup>37</sup> As shown in (35),  $IP_{[-PS\_LDD]}$  does not branch into any  $I'$ -adjoining rule which is critical for licensing the inner topic. That means the inner topic cannot be properly hosted by any phrase-structural rules. No formal solution can be produced for (45). As the formal grammar returns (45) as ungrammatical, this is in line with the generalisation about exhaustive subject control predicates, for which the displaced phrase must not appear inside the complement clause.

As a generalisation, the constraints in (46) are posited for the lexical entries of all exhaustive subject-control verbs.

$$(46) \quad (\uparrow XCOMP \text{ PS\_LDD}) = + \\ \text{CAT}((\uparrow XCOMP), \{IP_{[-PS\_LDD]}\})$$

Section 7.6 will discuss how to use a template, which is a formal device allowing commonalities to be represented succinctly, to capture the behaviour across all Chinese complementation verbs.

### 7.3 *Partial subject control (Patterns B and C)*

If a verb licenses partial subject control, the inner topic or focused phrase can either precede the partial-control verb or remain inside the embedded complement (Pattern B). When the displaced phrase precedes the partial-control verb, the embedded subject must be unexpressed (Pattern C). The bridging attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$  can be used to license the extraction of an inner topic or focused phrase into the matrix clause. However, no CAT predicate constraint is posited to impose any categorical requirement on its complement clause, unlike exhaustive subject control verbs.

The set of IP-associated rules, namely  $\{IP \rightarrow DP \ I', \ I' \rightarrow (DP) \ I', \ I' \rightarrow (I) \ VP, \ VP \rightarrow \dots V \dots IP \dots\}$  (with their functional annotations omitted here) are potentially recursive. There are two potential places for

<sup>37</sup> More accurately, the CAT predicate forces XCOMP to be associated with a set of nodes, one of which must contain  $IP_{[-PS\_LDD]}$ .

an I'-adjoining position to appear: higher or lower than the node of the matrix predicate (which occupies the V position). In other words, the displaced phrase can be structurally licensed either in the matrix clause or inside the complement clause. However, licensing the two potential structural positions alone is not sufficient. When the displaced phrase precedes the partial-control verb, the embedded subject must be unexpressed, suggesting the need for some additional constraint.

To demonstrate this, the lexical entry of the partial-control verb *jueding* 'decide' is presented in (47). The second line of its lexical entry involves an implicational constraint, which is conditioned by whether the embedded subject is realised in the c-structure. The formal definition of the function REALISED (Asudeh 2009, p. 111) is stated in (48). REALISED(*f*) requires c-structural realisation of f-structural elements.

(47) Lexical entry of *jueding* 'decide':

*jueding* 'decide' V ( $\uparrow$  PRED) = 'DECIDE < SUBJ, COMP >'  
 $\neg$ [REALISED( $\uparrow$  COMP SUBJ)]  
 $\Rightarrow$  [( $\uparrow$  COMP PS\_LDD) = +  
 $\wedge$ ( $\uparrow$  COMP SUBJ PRED) = 'PRO'  
 $\wedge$ ( $\uparrow$  SUBJ P\_CONTROL) = CONTROLLER  
 $\wedge$ ( $\uparrow$  COMP SUBJ P\_CONTROL) = CONTROLLEE]

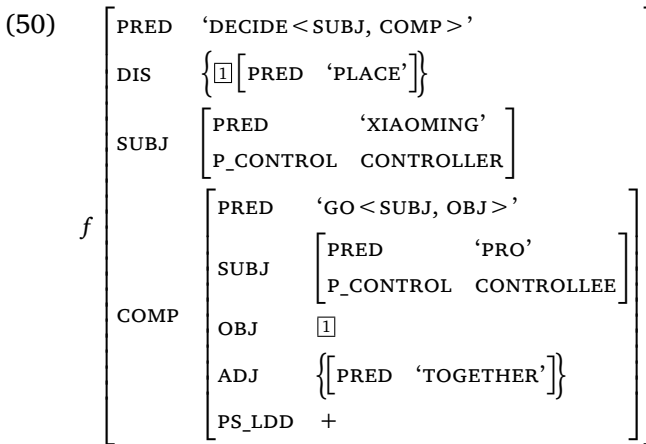
(48) REALISED function (Asudeh 2009, p. 111)

For any f-structure *f*, REALISED(*f*) is true if and only if  $\phi^{-1}(f) \neq \emptyset$ .

$\neg$ [REALISED(*f*)] requires the corresponding c-structural nodes to be unrealised. Only when the embedded subject is unrealised in the c-structure can the verb license partial control and assign the bridging attribute-value pair <PS\_LDD, +> to its clausal function COMP. The effect of this implicational constraint is manifested in (49).

(49) xiaoming<sub>i</sub> [zhe-ge difang] jueding { \*tamen | \*ta |  $\emptyset_{i+}$  }  
 Xiaoming this-CL place decide { they | 3SG |  $\emptyset$  }  
 yao yiqi qu  
 will together go  
 'Xiaoming decides to go to this place together.'

The required extraction path for the inner topic is ( $f$  COMP OBJ), where  $f$  is the  $f$ -structure of the matrix clause. The off-path constraint  $(\rightarrow \text{PS\_LDD}) =_c +$  imposed on the first clausal function COMP of the extraction path requires it to contain the attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$  in order for the out-of-complement-clause extraction to occur. However, when the embedded SUBJ is realised as *tamen* ‘they’ or *ta* ‘he/she’, the matrix predicate cannot assign the attribute-value pair due to the implicational constraint. Thus, these two configurations are rejected by the formal grammar. On the other hand, when the embedded SUBJ is unrealised, the implicational condition  $\neg[\text{REALISED}(\uparrow \text{COMP SUBJ})]$  is satisfied. The attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$  is assigned to the  $f$ -structure of COMP to license the extraction and the verb licenses partial control by assigning: (i) a pronominal value to its embedded subject; (ii) the attribute-value pair  $\langle \text{P\_CONTROL}, \text{CONTROLLER} \rangle$  to the matrix subject; and (iii) the attribute-value pair  $\langle \text{P\_CONTROL}, \text{CONTROLLEE} \rangle$  to the embedded subject (see Section 6). The well-formed  $c$ - and  $f$ -structure of (49) (with an unexpressed SUBJ) are presented in Figure 7 and in (50). From now on, we will skip the display of lexical information under the leaves of  $c$ -structural trees, reducing notational clutter.



Sentence (51) is another construction of *jueding* ‘decide’ with the inner topic residing in the complement clause. In contrast to 49), it is acceptable for the embedded SUBJ to be overt. Given the extrac-



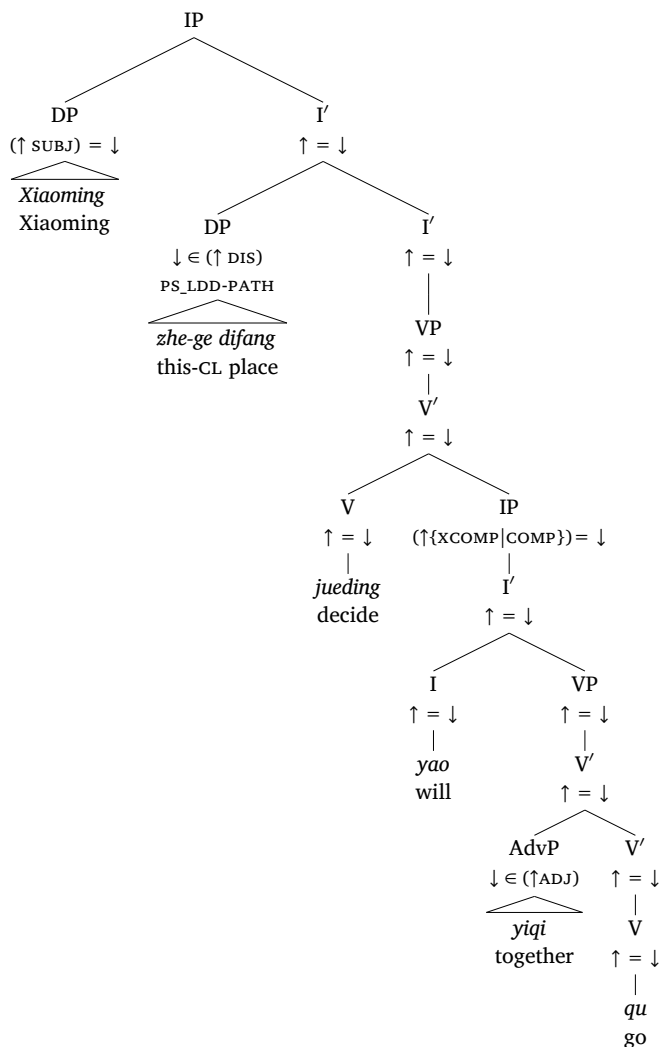


Figure 7: C-structure of (49)

tion path (g OBJ), there is no (first) clausal function which needs to be checked for the attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$ . Without any constraint violation, the formal grammar can successfully parse the construction. (52) shows what its f-structure looks like when its embedded SUBJ is overt and there is no partial control involved.

- (51) xiaowu<sub>i</sub> jue ding (tamen<sub>i+</sub>) [zhe-ge difang] yao yiqi qu  
 Xiaowu decide they this-CL place will together go  
 ‘Xiaowu decides that they will/to go to this place together.’

- (52) 
$$f \left[ \begin{array}{l} \text{PRED} \text{ 'DECIDE <SUBJ, COMP>} \\ \text{SUBJ} \left[ \text{PRED 'XIAOWU'} \right] \\ \text{COMP } g \left[ \begin{array}{l} \text{PRED 'GO <SUBJ, OBJ>} \\ \text{DIS } \left\{ \boxed{1} \left[ \text{PRED 'PLACE'} \right] \right\} \\ \text{SUBJ } \left[ \text{PRED 'PRO'} \right] \\ \text{OBJ } \boxed{1} \\ \text{ADJ } \left\{ \left[ \text{PRED 'TOGETHER'} \right] \right\} \end{array} \right] \end{array} \right]$$

As a generalisation, it is posited that all partial subject-control verbs contain the implicational constraint (53) in their lexical entries:

- (53)  $\neg[\text{REALISED}(\uparrow \text{COMP SUBJ})] \Rightarrow (\uparrow \text{COMP PS\_LDD}) = +$

## 7.4

*Non-control complementation (Pattern D)*

For a non-control complementation verb, its inner topic or focused phrase must remain inside the embedded complement. Non-control verbs and exhaustive-control verbs represent two ends of a spectrum regarding the capability of the matrix clause to host an inner topic or focused phrase. Earlier, it was discussed that the formal machinery for exhaustive-control verbs borrows insights from how LFG handles English bridge verbs. The lexically specified  $\langle \text{PS\_LDD}, + \rangle$  was devised as the bridging attribute-value pair to license a long-distance dependency relation that crosses the boundary of the embedded clause. The attribute  $\text{PS\_LDD}$  can be adopted for the  $f$ -structure of a non-control construction, but instead of the atomic value “+”, it is assigned the value “-”. The pair  $\langle \text{PS\_LDD}, - \rangle$  is lexically specified by a non-control predicate such as *xiangxin* ‘believe’ in (54). The extraction path  $\text{PS\_LDD-PATH}$  encoded in the  $I'$ -adjoining rule in (34) requires the first clausal function (if any) to contain the attribute-value pair  $\langle \text{PS\_LDD}, + \rangle$  via the off-path constraint  $(\rightarrow \text{PS\_LDD}) =_c +$ . Since the value of

PS\_LDD is now specified by *xiangxin* ‘believe’ to be “-”, it cannot satisfy the off-path constraining equation  $(\rightarrow \text{PS\_LDD}) =_c +$ . Therefore, a construction such as (55) is rejected by the formal grammar and its potential f-structure (56) is invalidated:

(54) Lexical entry of *xiangxin* ‘believe’:

*xiangxin* ‘believe’ V ( $\uparrow$  PRED) = ‘BELIEVE <SUBJ, COMP>’  
 ( $\uparrow$  COMP PS\_LDD) = -

(55) \*xiaoming [na-ben shu] xiangxin (ta) hui jinkuai  
 Xiaoming that-CL book believe 3SG will soon  
 wancheng  
 finish  
 ‘Xiaoming believes that he/she will finish that book soon.’

(56) Invalid f-structure:

[	PRED	‘BELIEVE <SUBJ, COMP>’	]
DIS	{	[1 [PRED ‘BOOK’]	]}
SUBJ	[	[PRED] ‘XIAOMING’	]
COMP	[	PRED ‘FINISH <SUBJ, OBJ>’	]
	SUBJ	[PRED ‘PRO’]	]
	OBJ	[1]	]
	PS_LDD	-	]

On the other hand, within-complement-clause extraction is permissible with the displaced phrase located in the post-subject position inside the complement clause. An example is given in (57) with its c- and f-structures presented in Figure 8 and in (58). Such a configuration is licensed: first, the off-path constraint  $(\rightarrow \text{PS\_LDD}) =_c +$  only applies to the first clausal function ever present; second, the path for within-complement-clause extraction ( $g$  OBJ) in (58) does not contain a clausal function. COMP in (58) corresponds to IP in Figure 8, whose set of associated rules includes the  $I'$ -adjoining rule for inner topicalisation and focus fronting.

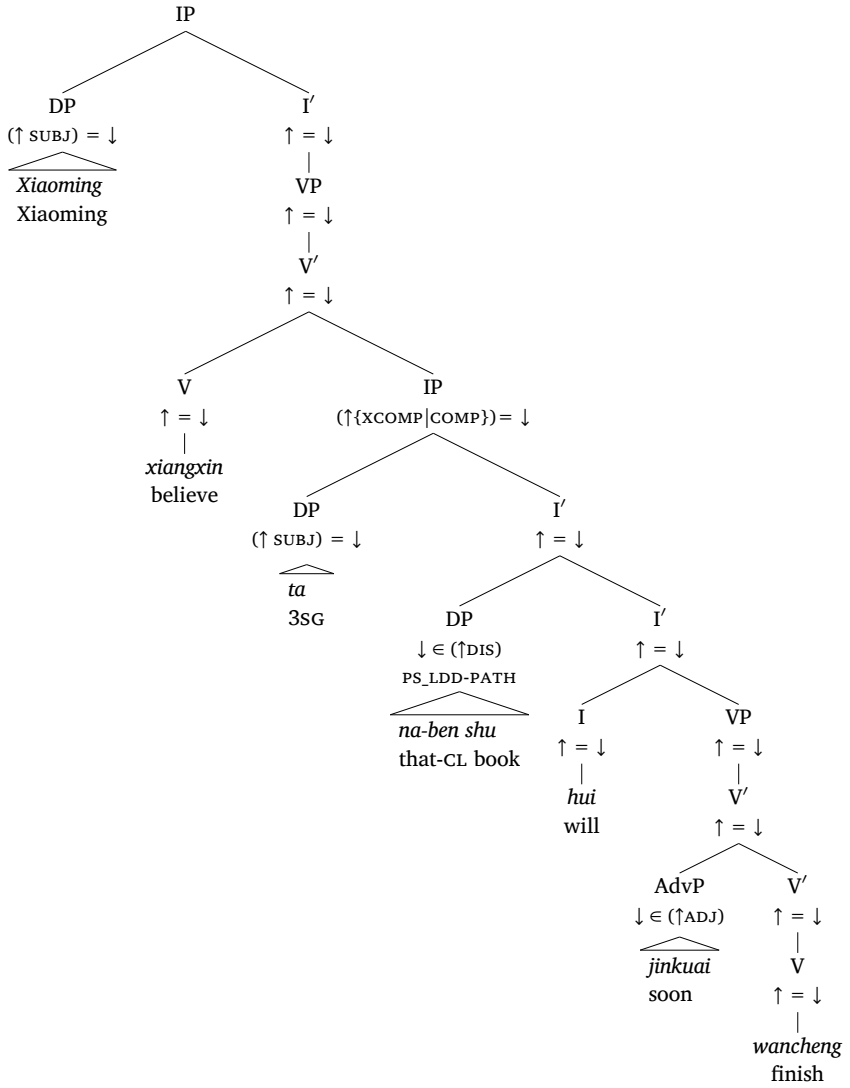


Figure 8: C-structure of (57)

- (57) xiaoming xiangxin (ta) [na-ben shu] hui jinkuai  
 Xiaoming believe 3SG that-CL book will soon  
 wancheng  
 finish  
 ‘Xiaoming believes that he/she will finish that book soon.’

- (58) 
$$f \left[ \begin{array}{l} \text{PRED} \quad \text{'BELIEVE <SUBJ, COMP>'} \\ \text{SUBJ} \quad \left[ \text{PRED} \quad \text{'XIAOMING'} \right] \\ \text{COMP} \quad g \left[ \begin{array}{l} \text{PRED} \quad \text{'FINISH <SUBJ, OBJ>'} \\ \text{DIS} \quad \left\{ \text{I} \left[ \text{PRED} \quad \text{'BOOK'} \right] \right\} \\ \text{SUBJ} \quad \left[ \text{PRED} \quad \text{'PRO'} \right] \\ \text{OBJ} \quad \text{I} \\ \text{PS-LDD} \quad - \end{array} \right] \end{array} \right]$$

As a generalisation, it is posited that all non-control verbs contain the constraint ( $\uparrow \text{COMP PS\_LDD}$ ) = – in their lexical entries.

*Object control (Pattern E)*

7.5

For an object-control verb, the inner topic or focused phrase must not precede the matrix-object controller, regardless of what control pattern the verb displays. Pre-theoretically, the matrix-object controller “blocks” the cross clausal boundary displacement, making the complement clause an unextractable island. While it may be tempting to associate some blocking device directly with the matrix-object controller, we argue that this treatment is dispreferred. For one thing, following the LFG analytical tradition (Section 6), the formal machinery here posits a lexically determined control mechanism. Thus, a grammatical function does not become a controller on its own merits but is accorded a controller status via the licensing constraints of the control verb. From this perspective, if a phenomenon appears to correlate with the identity of the controller, the entity which the phenomenon should ultimately be attributed to is the control verb. Therefore, we posit that for a construction with a matrix-object controller, its clausal function is assigned the attribute-value pair <PS\_LDD, –> by the object-control

verb, which is the same mechanism as that proposed for non-control verbs. As such, the lexicon regulates the displacement phenomena.

As an illustration, (59) is the lexical entry of *yuanliang* ‘forgive’ with a control equation and the constraint ( $\uparrow$  XCOMP PS\_LDD) = -. Sentence (60) is ill-formed and (61) is its invalid f-structure. In the extraction path ( $f$  XCOMP OBJ), the PS\_LDD feature in the f-structure of XCOMP has the value “-”, which renders the extraction impossible since the off-path constraint ( $\rightarrow$  PS\_LDD) = <sub>c</sub> + cannot be satisfied.

(59) Lexical entry of *yuanliang* ‘forgive’:

*yuanliang* ‘forgive’ V ( $\uparrow$  PRED) = ‘FORGIVE <SUBJ, OBJ, XCOMP>’  
 ( $\uparrow$  OBJ) = ( $\uparrow$  XCOMP SUBJ)  
 ( $\uparrow$  XCOMP PS\_LDD) = -

(60) \**xiaoming*<sub>i</sub> [*lian zhe-chang bisai*]        *dou yuanliang*  
 Xiaoming    even this-CL        competition PRT forgive  
*zhangsanj*  $\emptyset_{*i/j}$  *fangqi-le*  
 Zhangsan  $\emptyset$     give.up-PFV  
 ‘Xiaoming forgives Zhangsan to have given up even this competition.’

(61) Invalid f-structure:

PRED	‘FORGIVE <SUBJ, OBJ, COMP>’
DIS	$\left\{ \begin{array}{l} \boxed{1} \left[ \begin{array}{l} \text{PRED ‘COMPETITION’} \\ \text{SPEC [PRED ‘EVEN’]} \end{array} \right] \end{array} \right\}$
SUBJ	[PRED ‘XIAOMING’]
OBJ	$\boxed{2}$ [PRED ‘ZHANGSAN’]
XCOMP	$\left[ \begin{array}{l} \text{PRED ‘GIVE.UP <SUBJ, OBJ>’} \\ \text{SUBJ } \boxed{2} \\ \text{OBJ } \boxed{1} \\ \text{PS\_LDD -} \end{array} \right]$

Example (62) is a well-formed sentence displaying extraction within the complement clause, Figure 9 shows its c-structure, and (63) is its f-structure. An LFG syntactic tree does not need to obey binary

branching (Dalrymple *et al.* 2019, p. 98). The extraction path (g OBJ) is licensed since the off-path constraint ( $\rightarrow$  PS\_LDD) =  $c +$  in PS\_LDD-PATH only applies to the first clausal function which is absent in this case.

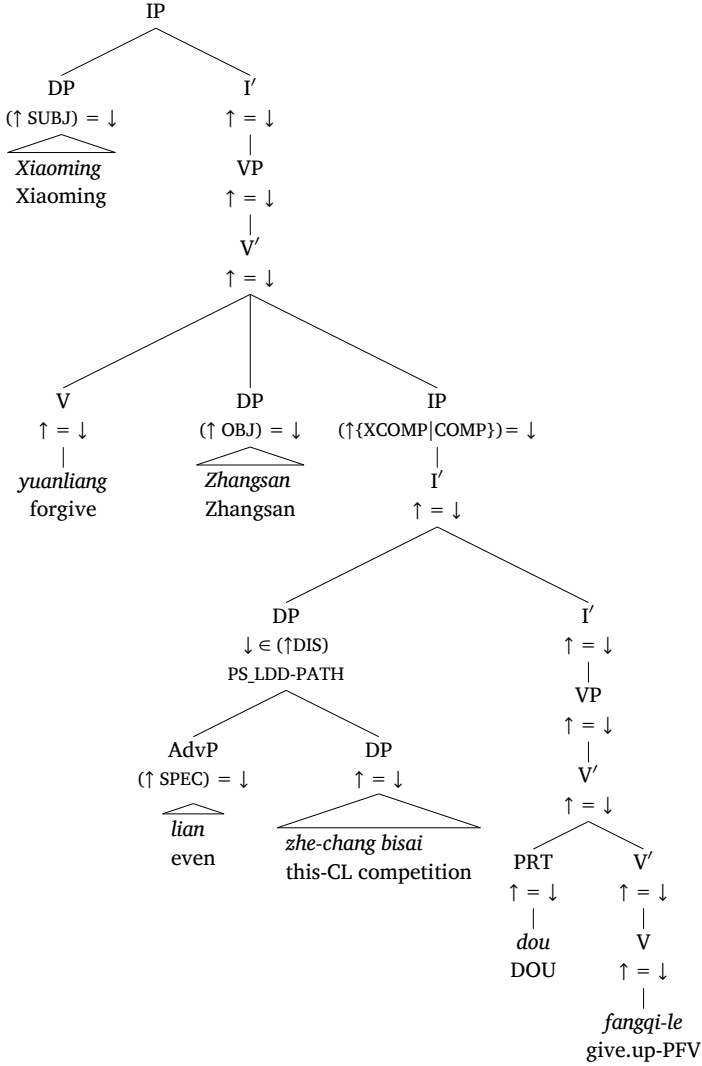


Figure 9: C-structure of (62)

- (62) xiaoming<sub>i</sub> yuanliang zhangsan<sub>j</sub> Ø<sub>\*i/j</sub> [lian zhe-chang  
 Xiaoming forgive Zhangsan Ø even this-CL  
 bisai] dou fangqi-le  
 competition PRT give.up-PFV  
 ‘Xiaoming forgives Zhangsan to have given up even this com-  
 petition.’

- (63) 
$$f \left[ \begin{array}{l} \text{PRED} \quad \text{'FORGIVE < SUBJ, OBJ, XCOMP >'} \\ \text{SUBJ} \quad \left[ \text{PRED 'XIAOMING'} \right] \\ \text{OBJ} \quad \boxed{1} \left[ \text{PRED 'ZHANGSAN'} \right] \\ \text{XCOMP} \quad g \left[ \begin{array}{l} \text{PRED} \quad \text{'GIVE.UP < SUBJ, OBJ >'} \\ \text{DIS} \quad \left\{ \boxed{2} \left[ \begin{array}{l} \text{PRED 'COMPETITION'} \\ \text{SPEC} \left[ \text{PRED 'EVEN'} \right] \end{array} \right] \right\} \\ \text{SUBJ} \quad \boxed{1} \\ \text{OBJ} \quad \boxed{2} \\ \text{PS\_LDD} \quad - \end{array} \right] \end{array} \right]$$

As a generalisation, it is posited that all object-control verbs contain the constraint (64) in their lexical entries:

- (64)  $(\uparrow \{ \text{XCOMP} | \text{COMP} \} \text{PS\_LDD}) = -$

7.6

Template for complementation verbs

In LFG, it is possible to capture commonalities between lexical entries via a formal device known as a “template”, which allows “commonalities between lexical entries to be represented succinctly and linguistic generalizations to be encoded in a theoretically motivated manner” (Dalrymple *et al.* 2019, p. 234). We posit that all Chinese complementation verbs share the template VCOMPINTOPFOCFRONT in (65), which encodes correlations among control properties, inner topicalisation, and focus fronting:



- (65) VCOMPINTOPFOCFRONT  $\equiv$
- $$\left\{ \begin{array}{l} \{ (\uparrow \text{OBJ}) = (\uparrow \text{XCOMP SUBJ}) \mid (\uparrow \text{OBJ P\_CONTROL}) = \text{CONTROLLER} \} \\ \Rightarrow (\uparrow \{ \text{XCOMP} \mid \text{COMP} \} \text{PS\_LDD}) = - \\ \mid (\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ}) \\ \Rightarrow (\uparrow \text{XCOMP PS\_LDD}) = + \wedge \text{CAT}((\uparrow \text{XCOMP}), \{ \text{IP}_{[-\text{PS\_LDD}]} \}) \\ \mid (\uparrow \text{SUBJ P\_CONTROL}) = \text{CONTROLLER} \\ \Rightarrow (\uparrow \text{COMP PS\_LDD}) = + \\ \mid \neg (\uparrow \text{XCOMP}) \wedge \neg (\uparrow \text{COMP SUBJ P\_CONTROL}) \\ \Rightarrow (\uparrow \text{COMP PS\_LDD}) = - \\ \} \end{array} \right.$$

The template VCOMPINTOPFOCFRONT contains four (broad) disjunctive options. The first option targets object-control verbs, which are featured by possessing either the functional-control equation  $(\uparrow \text{OBJ}) = (\uparrow \text{XCOMP SUBJ})$  or one of the constraints for partial control  $(\uparrow \text{OBJ P\_CONTROL}) = \text{CONTROLLER}$ . The second option targets exhaustive subject-control verbs, which are characterised by the functional-control equation  $(\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ})$ . The third option targets partial subject-control verbs, which contain the constraint  $(\uparrow \text{SUBJ P\_CONTROL}) = \text{CONTROLLER}$  for encoding its controller function. The fourth option targets non-control complementation verbs, which neither subcategorise for XCOMP nor assign the attribute P\_CONTROL to the f-structure of its complement-clause subject. We can rewrite the lexical entries of *shefa* ‘try’ (exhaustive subject-control), *jueding* ‘decide’ (partial subject-control), *xiangxin* ‘believe’ (non-control), and *yuanliang* ‘forgive’ (object-control) as follows. All of them share the same template VCOMPINTOPFOCFRONT, which interacts with other constraints in the lexical entry to generate the desired displacement patterns:

- (66)
- |                         |  |
|-------------------------|--|
| <i>shefa</i> ‘try’      | V $(\uparrow \text{PRED}) = \text{‘TRY<SUBJ, XCOMP>’}$<br>$(\uparrow \text{SUBJ}) = (\uparrow \text{XCOMP SUBJ})$<br>@VCOMPINTOPFOCFRONT   |
| <i>jueding</i> ‘decide’ | V $(\uparrow \text{PRED}) = \text{‘DECIDE<SUBJ, COMP>’}$<br>$\neg[\text{REALISED}(\uparrow \text{COMP SUBJ})]$<br>$\Rightarrow [(\uparrow \text{COMP SUBJ PRED}) = \text{‘PRO’}]$<br>$\wedge(\uparrow \text{SUBJ P\_CONTROL}) = \text{CONTROLLER}$<br>$\wedge(\uparrow \text{COMP SUBJ P\_CONTROL}) = \text{CONTROLLEE}]$<br>@VCOMPINTOPFOCFRONT |

- xiangxin* ‘believe’ V (↑ PRED) = ‘BELIEVE < SUBJ, COMP >’  
@VCOMPINTOPFOCFRONT
- yuanyang* ‘forgive’ V (↑ PRED) = ‘FORGIVE < SUBJ, OBJ, XCOMP >’  
(↑ OBJ) = (↑ XCOMP SUBJ)  
@VCOMPINTOPFOCFRONT

## 7.7

*Complex embedding and extraction paths*

It was mentioned earlier that there is a competing version of the formula PS\_LDD-PATH governing possible extraction paths:

- (67) a. Correct version of PS\_LDD-PATH  
 $(\uparrow (\{XCOMP|COMP\} \{XCOMP|COMP\}^*) OBJ) = \downarrow$   
 $(\rightarrow PS\_LDD) =_c +$
- b. Competing but incorrect version of PS\_LDD-PATH  
 $(\uparrow \{XCOMP|COMP\}^* OBJ) = \downarrow$   
 $(\rightarrow PS\_LDD) =_c +$

To understand why (67b) makes wrong predictions, one needs to turn to complex embedding, involving two or more clause-embedding verbs. (68) contains complex-embedding constructions of five clausal levels. The first four levels are headed by complementation predicates – *jueding* ‘decide’, *quan* ‘try to persuade’, *xiangbanfa* ‘try/strive’, and *shou* ‘say’. Among them, *jueding* ‘decide’ and *xiangbanfa* ‘try/strive’ assign the attribute-value pair <PS\_LDD, +> to their respective complement clause, whereas *quan* ‘try to persuade’ and *shou* ‘say’ assign <PS\_LDD, ->. (68) and Table 15 examine the logically possible places for *zhe-jian shiqing* ‘this thing’ when it is used as an inner topic. Table 15 has boxed those functions that have received <PS\_LDD, +>. <sup>38</sup>

- (68) a. *xiaoming jueding* Ø *quan* *xiaomei* Ø  
 Xiaoming decide Ø try.to.persuade Xiaomei Ø  
*xiangbanfa* Ø *gen pengyou shuo* Ø [zhe-jian  
 try Ø to friend say Ø this-CL  
*shiqing*] *meiyou zuo-guo*  
 thing not do-EXP  
 ‘Xiaoming decides to persuade Xiaomei to try to say to friends that (somebody) has not done this thing.’

<sup>38</sup>In Table 15, ↑ refers to the f-structure immediately enclosing the inner-topic function DIS.

- b. \*xiaoming jueding  $\emptyset$  quan                      xiaomei  $\emptyset$   
Xiaoming decide  $\emptyset$  try.to.persuade Xiaomei  $\emptyset$   
xiangbanfa  $\emptyset$  [zhe-jian shiqing] gen pengyou shuo  
try  $\emptyset$  this-CL thing to friend say  
 $\emptyset$  meiyou zuo-guo  
 $\emptyset$  not do-EXP  
'Xiaoming decides to persuade Xiaomei to try to say to  
friends that (somebody) has not done this thing.'
- c. xiaoming jueding  $\emptyset$  quan                      Xiaomei  $\emptyset$  [zhe-jian  
Xiaoming decide  $\emptyset$  try.to.persuade Xiaomei  $\emptyset$  this-CL  
shiqing] xiangbanfa  $\emptyset$  gen pengyou shuo  $\emptyset$  meiyou  
thing try  $\emptyset$  to friend say  $\emptyset$  not  
zuo-guo  
do-EXP  
'Xiaoming decides to persuade Xiaomei to try to say to  
friends that (somebody) has not done this thing.'
- d. \*xiaoming jueding  $\emptyset$  [zhe-jian shiqing] quan  
Xiaoming decide  $\emptyset$  this-CL thing try.to.persuade  
xiaomei  $\emptyset$  xiangbanfa  $\emptyset$  gen pengyou shuo  $\emptyset$   
Xiaomei  $\emptyset$  try  $\emptyset$  to friend say  $\emptyset$   
meiyou zuo-guo  
not do-EXP  
'Xiaoming decides to persuade Xiaomei to try to say to  
friends that (somebody) has not done this thing.'
- e. xiaoming [zhe-jian shiqing] jueding  $\emptyset$  quan  
Xiaoming this-CL thing decide  $\emptyset$  try.to.persuade  
xiaomei  $\emptyset$  xiangbanfa  $\emptyset$  gen pengyou shuo  $\emptyset$   
Xiaomei  $\emptyset$  try  $\emptyset$  to friend say  $\emptyset$   
meiyou zuo-guo  
not do-EXP  
'Xiaoming decides to persuade Xiaomei to try to say to  
friends that (somebody) has not done this thing.'

Both versions of the PS\_LDD-PATH formula in (67a) and (67b) give the correct predictions about the acceptability of (68a) and (68b). However, only (67a) makes correct predictions about the acceptabil-

Table 15:  
Displacement  
patterns,  
extraction paths,  
and acceptability

Displacement pattern	Extraction path	Acceptability
(68a)	(↑ OBJ)	Acceptable
(68b)	(↑ COMP OBJ)	Unacceptable
(68c)	(↑ <span style="border: 1px solid black; padding: 0 2px;">COMP</span> COMP OBJ)	Acceptable
(68d)	(↑ XCOMP <span style="border: 1px solid black; padding: 0 2px;">COMP</span> COMP OBJ)	Unacceptable
(68e)	(↑ <span style="border: 1px solid black; padding: 0 2px;">COMP</span> XCOMP <span style="border: 1px solid black; padding: 0 2px;">COMP</span> COMP OBJ)	Acceptable

ity of all the sentences. If an extraction path contains more than one clausal function, only the first clausal function is required to contain <PS\_LDD, + >. From another perspective, whether it is possible for a displaced phrase to be extracted out of a complement clause depends on the licensing properties of the complementation verb that is on the same clausal level (in the f-structure) as the DIS function borne by the displaced phrase.

## 8 COMPUTATIONAL IMPLEMENTATION FOR CONSTRAINT TESTING

Section 7 has presented a theoretical LFG analysis. In order to safeguard the formal accuracy of the constraints and oversee their complex interaction – particularly, the interaction among control, complementation, inner topicalisation and focus fronting – we have computationally implemented the analysis using a grammar-engineering tool – Xerox Linguistic Environment (XLE; Crouch *et al.* 2011).<sup>39</sup> The results of computational testing are included in Appendix B, providing evidence that our proposed constraints are not only theoretically possible but also computationally implementable.<sup>40</sup>

<sup>39</sup>XLE has been used in the Parallel Grammar Project (ParGram; <https://pargram.w.uib.no/>; Sulger *et al.* 2013) to develop cross-linguistic computational grammars.

<sup>40</sup>For further information about the computational implementation of grammatical formalisms, one may refer to two special issues of the *Journal of Language Modelling*: Volume 10, Number 1, the 2022 issue on the interaction between for-

## CONCLUSION

9

This paper examined the empirical data of inner topicalisation and focus fronting, focusing on their interaction with control and complementation. Our discussion led to five empirical generalisations, which were further tested using acceptability-judgment tasks on a subset of complementation verbs. We have proposed a formal lexicalist analysis to capture the correlational relationships, which differs from existing restructuring analyses. Our non-movement proposal uses LFG's formalism of long-distance dependency, where displacement is not contingent on the size of the embedded clause. We argue that this approach better captures the empirical facts of inner topicalisation and focus fronting than restructuring accounts. Given the computational rigour of LFG, we have implemented our analysis using XLE. The computational implementation provides further evidence about the formal accuracy of our proposed constraints.

---

mal and computational linguistics; and Volume 3, Number 1, the 2015 issue on methodologies for grammar engineering. Computationally implemented grammars allow linguists to test analyses and keep track of the interaction between different parts of the grammar, besides any other technological applications for which they can be used. See, e.g., Forst and King 2023, Zamaraeva *et al.* 2022, Bernard and Winterstein 2022, Duchier and Parmentier 2015, Müller 2015, and Bender 2008.

## APPENDICES

### A SAMPLE STIMULI

There were in total five acceptability-judgment tasks. Each acceptability-judgment task contained four conditions. Each condition had four lexicalisations. The test sentences were distributed in a Latin square design for counterbalancing such that no sentences in a list were variants of each other.

For example, in Task 1, there were four conditions, and each condition contained four lexicalisations describing the following scenarios: (i) end-of-term exam, (ii) mathematical question, (iii) Olympic event, (iv) washing dishes. For every scenario, there were four minimal variants distributed across the four conditions. In this Appendix, we will demonstrate one lexicalisation (out of four lexicalisations) for each condition. English glosses are added in this Appendix for illustrative purposes, but the stimuli were presented only in written Chinese to the participants.

#### A.1 *Acceptability-judgment task 1*

*Condition A: Crossing  $V_m$  + Focus Fronting (Exhaustive Control)*

- (1) *Context: Tomorrow is the day of the important end-of-term exam.*  
xiaoding [lian ruci zhongyao-de qimo kaoshi] dou  
Xiaoding even so important-DE end.of.term exam PRT  
shefa zhao jikou bu canjia  
try find excuse not take.part  
'Xiaoding tries to find an excuse not to take part in even such an important end-of-term exam.'

*Condition B: Not crossing  $V_m$  + Focus Fronting (Exhaustive Control)*

- (2) *Context: This is a challenging mathematical question.*  
xiaohong shefa [lian zhe-dao name shenao-de shuxue  
Xiaoding try even this-CL so challenging-DE maths  
nanti] dou jieju  
question PRT solve  
'Xiaoding tries to solve even such a challenging mathematical question.'

Condition C: Crossing  $V_m$  + Inner Topicalisation (Exhaustive Control)

(3) Context: This Olympic event is intense.

yuehan [zhe-chang bisai] neng shefa shengchu  
John this-CL competition able try win  
'John tries to win this competition.'

Condition D: Not crossing  $V_m$  + Inner Topicalisation (Exhaustive Control)

(4) Context: Washing dishes is not a difficult task.

keshi lisi shefa [zhe-zhong shiqing] jiao gei bieren qu zuo  
but Lisi try this-CL task pass to others go do  
'Lisi tries to pass on this task to others.'

Acceptability-judgment task 2

A.2

Condition A: Crossing  $V_m$  + Focus Fronting (Partial Control)

(5) Context: Xiaoli always handles everything himself.

xiaoli [lian ruci suosui-de shiqing] dou xiangyao ziji  
Xiaoli even so trivial-DE matter PRT want SELF  
chuli  
handle  
'Xiaoli wants to handle even such a trivial matter by himself.'

Condition B: Not crossing  $V_m$  + Focus Fronting (Partial Control)

(6) Context: This report is especially long.

xiaoming xiangyao [lian zhe-pian tebie zhang-de  
Xiaoming want even this-CL especially long-DE  
baogao] dou jinkuai xiewan  
report PRT soon finish  
'Xiaoming wants to finish even such a long report soon.'

Condition C: Crossing  $V_m$  + Inner Topicalisation (Partial Control)

(7) Context: Buddha's Temptation is a highly challenging dish.

xiaowang [zhe-dao cai] xiangyao shunli zuochu  
Xiaowang this-CL dish want successfully make  
'Xiaowang wants to make this dish successfully.'

*Condition D: Not crossing  $V_m$  + Inner Topicalisation (Partial Control)*

- (8) *Context: This movie is very difficult to grasp.*  
xiaodong xiangyao [zhe-bu dianying] kandedong  
Xiaodong want this-CL movie understand  
'Xiaodong wants to understand this movie.'

A.3

*Acceptability-judgment task 3*

*Condition A: SUBJ unexpressed + Focus Fronting (Partial Control)*

- (9) *Context: The boss is always very efficient.*  
lingdao [lian ruci jianju-de renwu] dou jueding yao zai  
boss even so difficult-DE task PRT decide need at  
mingtian nei wancheng  
tomorrow within finish  
'The boss decides to finish even such a difficult task by the end of tomorrow.'

*Condition B: SUBJ expressed + Focus Fronting (Partial Control)*

- (10) *Context: Xiaoming is a very smart student.*  
xiaoming [lian name nanzuo-de gongke] dou jueding  
Xiaoming even such difficult-DE assignment PRT decide  
ta yao zai yitian nei tijiao  
3SG need at one.day within submit  
'Xiaoming decides to submit even such a difficult assignment within a day.'

*Condition C: SUBJ unexpressed + Inner Topicalisation (Partial Control)*

- (11) *Context: Xiaoxiu has announced her retirement from the film industry. Will she still take this movie?*  
xiaoxiu [zhe-bu dianying] jueding bu hui jie  
Xiaoxiu this-CL movie decide not will take  
'Xiaoxiu decides not to take this movie.'



Condition D: SUBJ expressed + Inner Topicalisation (Partial Control)

- (12) Context: Xiaogang does not like people sending him gifts. Will he accept this gift?

xiaogang [zhe-fen liwu] jueding ta bu hui shouxia  
Xiaogang this-CL gift decide 3SG not will accept  
'Xiaogang decides not to accept this gift.'

Acceptability-judgment task 4

A.4

Condition A: Crossing  $V_m$  + Focus Fronting (Non-control)

- (13) Context: Xiaowang is good at imitating sounds.

xiaowang [lian dongwu-de shengyin] dou shuo-guo  
Xiaowang even animal-DE sound PRT say-EXP  
nengguo mofang  
can imitate  
'Xiaowang has said (he) can imitate even animal sounds.'

Condition B: Not crossing  $V_m$  + Focus Fronting (Non-control)

- (14) Context: Xiaojie is an excellent writer.

xiaojie shuo-guo [lian zhe-ben changpian xiaoshuo] dou  
Xiaojie say-EXP even this-CL long novel PRT  
neng zai yi-ge yue nei xiewan  
can at one-CL month within finish  
'Xiaojie has said (he) can finish even such a long novel within a month.'

Condition C: Crossing  $V_m$  + Inner Topicalisation (Non-control)

- (15) Context: Xiaojian is good at designing computer games.

xiaojian [zhe-kuan diannaoyouxi] shuo-guo neng sheji  
Xiaojian this-CL computer game say-EXP can design  
hao  
well  
'Xiaojian has said (he) can design this computer game well.'

*Condition D: Not crossing  $V_m$  + Inner Topicalisation (Non-control)*

- (16) *Context: Does Xiaonan want to visit this country?*  
xiaonan shuo-guo [zhe-ge guojia] bu hui qu  
Xiaonan say-EXP this-CL country not will go  
'Xiaonan has said (he) will not go to this country.'

A.5

*Acceptability-judgment task 5*

*Condition A: Crossing  $OBJ_m$  controller + Focus Fronting*

- (17) *Context: This book is very difficult to understand.*  
xiaoming [lian zhe-ben ruci shenao-de shu] dou  
Xiaoming even this-CL so difficult-DE book PRT  
shuifu-le xiaomei yao hao hao du  
persuade-PFV Xiaomei need.to properly read  
'Xiaoming has persuaded Xiaomei to read even such a difficult  
book properly.'

*Condition B: Not crossing  $OBJ_m$  controller + Focus Fronting*

- (18) *Context: There will be an important competition tomorrow.*  
mama shuifu-le zhangsan [lian zhe-chang ruci  
mum persuade-PFV Zhangsan even this-CL so  
zhongyao-de bisai] dou dei fangqi  
important-DE competition PRT need.to give.up  
'Mum has persuaded Zhangsan to give up even such an impor-  
tant competition.'

*Condition C: Crossing  $OBJ_m$  controller + Inner Topicalisation*

- (19) *Context: This oil painting is very expensive.*  
chen xiaojie [zhe-fu youhua] shuofu-le ceng  
Chen Miss this-CL oil.painting persuade-PFV Ceng  
xiansheng yao goumai  
Mr. need.to buy  
'Miss Chen has persuaded Mr. Ceng to buy this oil painting.'

Condition D: Not crossing OBJ<sub>m</sub> controller + Inner Topicalisation

- (20) Context: *This traditional musical instrument is very hard to learn.*  
didi           shuifu-le       gege           [zhe-jian  
young.brother persuade-PFV elder.brother this-CL  
chuantong yueqi]   yao   qu xue  
traditional instrument need.to go learn  
'The younger brother has persuaded the elder brother to learn  
this traditional instrument.'

COMPUTATIONAL IMPLEMENTATION  
AND GRAMMAR TESTING ON XLE

B

To safeguard the formal accuracy of our constraints and oversee their complex interaction, we have computationally tested our theoretical analysis by implementing it on the grammar-engineering tool Xerox Linguistic Environment (XLE; Crouch *et al.* 2011).<sup>41</sup> We present some important constraints in our computational grammar, which has incorporated those constraints discussed in Sections 7.1–7.5. Here, the constraints are stated in a way that follows XLE's computational requirements. For more information, please refer to the XLE documentation (Crouch *et al.* 2011). The following are c-structural rules,

---

<sup>41</sup> As pointed out by Bender (2008, p. 16): "Grammar engineering is the process of creating machine-readable implementations of formal grammars... Computerized implementations of their grammars allow linguists to more efficiently and effectively test hypotheses... Languages are made up of many subsystems with complex interactions. Linguists generally focus on just one subsystem at a time, yet the predictions of any particular analysis cannot be calculated independently of the interacting subsystems. With implemented grammars, the computer can track the effects of all aspects of the implementation while the linguist focuses on developing just one."

lexical entries, and templates.<sup>42</sup> As a recap, *shefa* ‘try’ is an exhaustive-control verb. Both *dasuan* ‘intend’ and *jueding* ‘decide’ are partial-control verbs; *jueding* ‘decide’ allows its embedded subject to be optionally expressed but *dasuan* ‘intend’ does not. *Xiangxin* ‘believe’ is a non-control verb and *yuanliang* ‘forgive’ is an object-control verb. Note that we have defined the if-then logical operation using a parametrised template and we have used the CAT predicate to help define relations involving the inverse correspondence  $\phi^{-1}$ . The epsilon  $\epsilon$  is used on XLE to designate an empty string, which will not be displayed in the c-structure.

## B.1

*C-structural rules (XLE)*

```

IP --> (DP: (^ SUBJ)=!)
        I': ^=!.

I' --> {DP: !$ (^DIS)
        @(PS_LDD-PATH);
        I': ^=!
        |(I)
        VP: ^=!
        }.

VP --> V': ^=!.

V' --> {PRT: ^=!;
        V': ^=!
        |AdvP: !$(^ADJUNCT);
        V': ^=!
        |V: ^=!;
        (DP: (^OBJ)=!)
        ({IP: (^{XCOMP|COMP})=!
         |IP[-PS_LDD]: (^XCOMP)=!
         })
        }.

```

---

<sup>42</sup>Since the internal structure of Chinese noun phrases is not our focus, our computational grammar tends to simplify it. For example, *zhe-xiang* ‘this-CL’ is represented as one demonstrative in the c-structure.

IP[-PS\_LDD] --> I'[-PS\_LDD]: ^=!

I'[-PS\_LDD] --> (I)  
VP: ^=!

DP --> {(D)  
AP\*: ! \$ (^ADJUNCT);  
N: ^=!;  
|AdvP: (^SPEC)=!;  
DP: ^=!  
}.

*(Parametrised) templates (XLE)*

B.2

PS\_LDD-PATH = (^({XCOMP:(->PS\_LDD)=c+;  
|COMP:(->PS\_LDD)=c+;}{XCOMP|COMP}\*})OBJ)=!

EC-SUBJ(P) = (^PRED) = 'P<(^SUBJ)(^XCOMP)>'  
(^SUBJ) = (^XCOMP SUBJ).

EC-OBJ(P) = (^PRED) = 'P<(^SUBJ)(^OBJ)(^XCOMP)>'  
(^OBJ) = (^XCOMP SUBJ).

PC-SUBJ(P) = (^PRED) = 'P<(^SUBJ)(^COMP)>'  
(^COMP SUBJ PRED) = 'PRO'  
(^SUBJ P\_CONTROL) = CONTROLLER  
(^COMP SUBJ P\_CONTROL) = CONTROLLEE.

PC-optional-SUBJ(P) = (^PRED) = 'P<(^SUBJ)(^COMP)>'  
@(IF @(CAT(^COMP SUBJ PRED) e) @PC-PS\_LDD-optional).

PC-PS\_LDD-optional = @(PC-PS\_LDD)  
(^COMP SUBJ PRED) = 'PRO'  
(^SUBJ P\_CONTROL) = CONTROLLER  
(^COMP SUBJ P\_CONTROL) = CONTROLLEE.

VCOMP(P) = (^PRED) = 'P<(^SUBJ)(^COMP)>'.

IF(P Q) = {~P |~~P Q}.

EC-PS\_LDD = (^XCOMP PS\_LDD) = +  
          @(CAT (^XCOMP) IP[-PS\_LDD]).

PC-PS\_LDD = (^COMP PS\_LDD) = +.

OBJ-PS\_LDD = (^{XCOMP|COMP} PS\_LDD) = -.

NC-PS\_LDD = (^COMP PS\_LDD) = -.

### B.3

#### *Lexical entries (XLE)*

shefa V \* @(EC-SUBJ try)  
          @(EC-PS\_LDD).

dasuan V \* @(PC-SUBJ intend)  
          @(PC-PS\_LDD).

jueding V \* @(PC-optional-SUBJ decide).

yuanliang V \* @(EC-OBJ forgive)  
          @(OBJ-PS\_LDD).

xiangxin V \* @(VCOMP believe)  
          @(NC-PS\_LDD).

### B.4

#### *Test cases*

We now turn to the test suite, which contains a series of sentences fed to the computational grammar for constraint testing. All parsing results are in line with our predictions discussed in Section 7. In what follows, we will illustrate a set of test cases.<sup>43</sup> For brevity, we will only present inner topicalisation in this Appendix. The same results

---

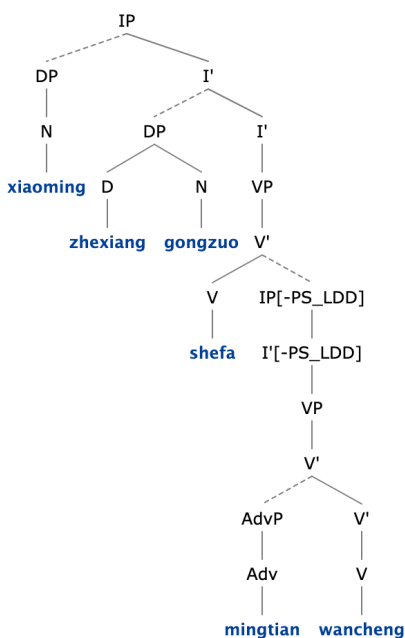
<sup>43</sup>Our grammar fragment was loaded to the XLE-web interface developed at the University of Konstanz (<https://ling.sprachwiss.uni-konstanz.de/pages/xle/iness.html>), which is based on the XLE Web interface on INESS (Rosén *et al.* 2012).

have been obtained for focus fronting with regard to the position of the displaced phrase. We will also present some complex-embedding test cases.

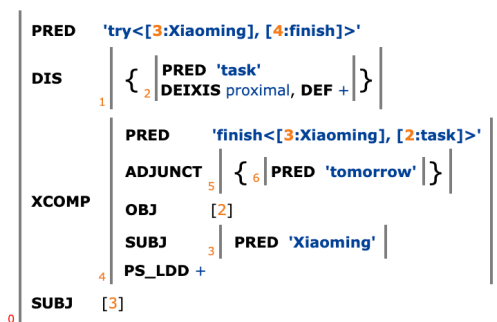
**Test case 1: Exhaustive subject control with the inner topic crossing the control verb**

- (1) xiaoming zhexiang gongzuo shefa mingtian wancheng  
 xiaoming this-CL task try tomorrow finish  
 ‘Xiaoming tries to finish this task tomorrow.’

**C-structure**



**F-structure**



**Test case 2: Exhaustive subject control with the inner topic residing in the complement clause**

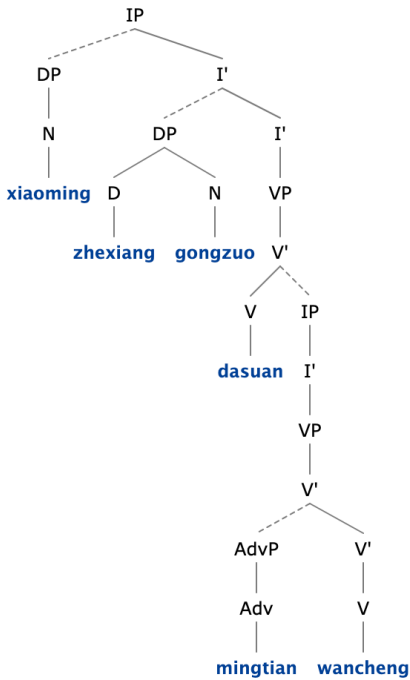
- (2) xiaoming shefa zhe-xiang gongzuo mingtian wancheng  
 Xiaoming try this-CL task tomorrow finish  
 ‘Xiaoming tries to finish this task tomorrow.’

No formal solution could be produced by our grammar fragment for test case 2.

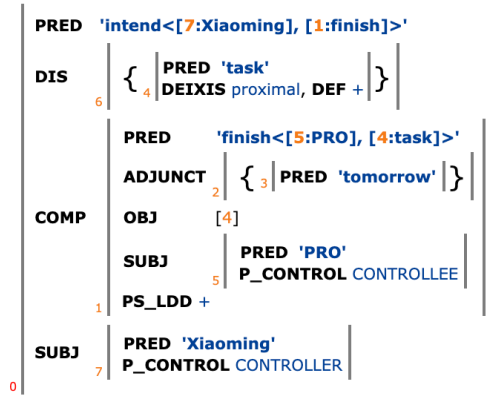
**Test cases 3–4: Partial subject control with the inner topic crossing the control verb (unexpressed embedded subj)**

- (3) xiaoming zhe-xiang gongzuo dasuan mingtian wancheng  
 Xiaoming this-CL task intend tomorrow finish  
 ‘Xiaoming intends to finish this task tomorrow.’

**C-structure**



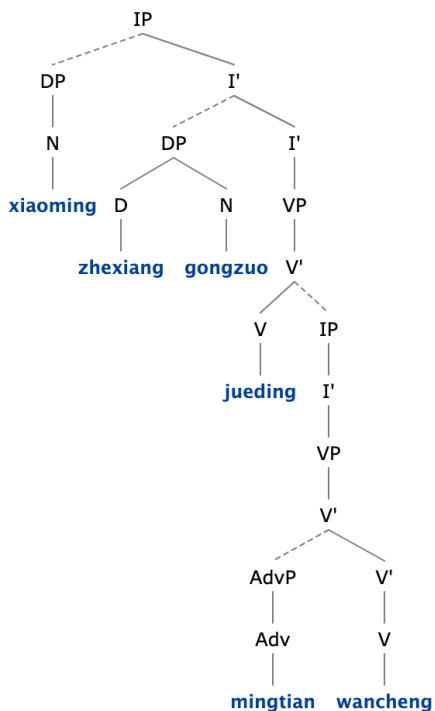
**F-structure**



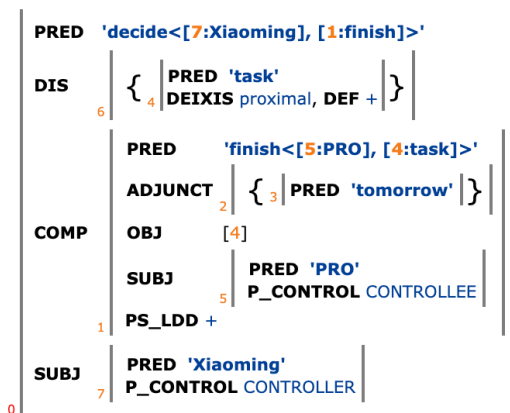


- (4) xiaoming zhe-xiang gongzuo jueding mingtian wancheng  
 Xiaoming this-CL task decide tomorrow finish  
 'Xiaoming decides to finish this task tomorrow.'

### C-structure



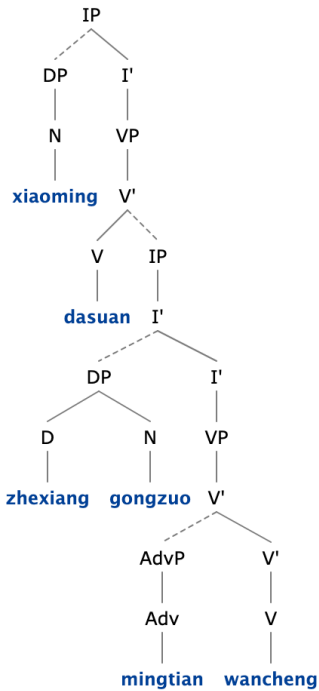
### F-structure



**Test cases 5–6: Partial subject control with the inner topic residing in the complement clause (unexpressed embedded subject)**

- (5) xiaoming dasuan zhe-xiang gongzuo mingtian wancheng  
 Xiaoming intend this-CL task tomorrow finish  
 ‘Xiaoming intends to finish this task tomorrow.’

**C-structure**

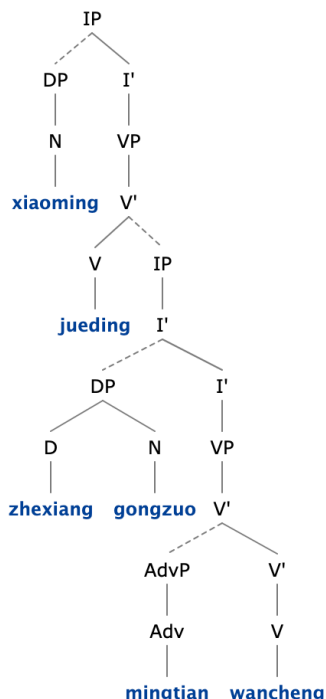


**F-structure**

<b>PRED</b>	'intend<[7:Xiaoming], [1:finish]>'								
<b>DIS</b>	'finish<[6:PRO], [5:task]>'								
<b>COMP</b>	<table border="1"> <tr> <td><b>ADJUNCT</b></td> <td>{ 3   PRED 'tomorrow' }</td> </tr> <tr> <td><b>OBJ</b></td> <td>[5]</td> </tr> <tr> <td><b>SUBJ</b></td> <td>PRED 'PRO' P_CONTROL CONTROLLEE</td> </tr> <tr> <td><b>PS_LDD +</b></td> <td></td> </tr> </table>	<b>ADJUNCT</b>	{ 3   PRED 'tomorrow' }	<b>OBJ</b>	[5]	<b>SUBJ</b>	PRED 'PRO' P_CONTROL CONTROLLEE	<b>PS_LDD +</b>	
<b>ADJUNCT</b>	{ 3   PRED 'tomorrow' }								
<b>OBJ</b>	[5]								
<b>SUBJ</b>	PRED 'PRO' P_CONTROL CONTROLLEE								
<b>PS_LDD +</b>									
<b>SUBJ</b>	PRED 'Xiaoming' P_CONTROL CONTROLLER								
	0								

- (6) xiaoming jueding zhe-xiang gongzuo mingtian wancheng  
 Xiaoming decide this-CL task tomorrow finish  
 ‘Xiaoming decides to finish this task tomorrow.’

**C-structure**



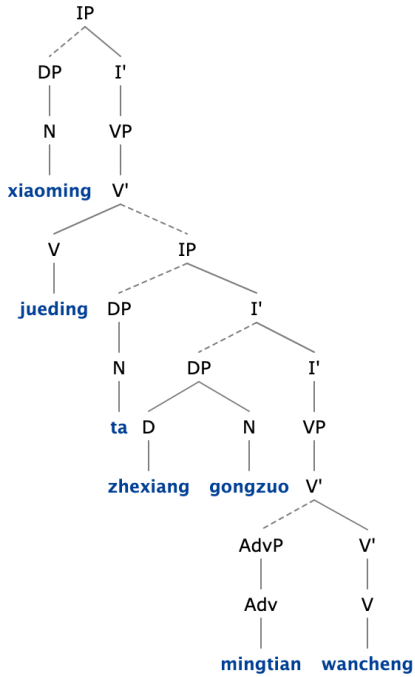
**F-structure**

<b>PRED</b>	'decide<[8:Xiaoming], [1:finish]>'
<b>PRED</b>	'finish<[6:PRO], [5:task]>'
<b>DIS</b>	{ PRED 'task' DEIXIS proximal, DEF + }
<b>COMP</b>	ADJUNCT { PRED 'tomorrow' }
<b>OBJ</b>	[5]
<b>SUBJ</b>	PRED 'PRO' P_CONTROL CONTROLLEE
<b>PS_LDD +</b>	[6]
<b>SUBJ</b>	PRED 'Xiaoming' P_CONTROL CONTROLLER
	[8]

Test case 7: Partial subject control with the inner topic residing in the complement clause (expressed embedded subject)

- (7) xiaoming jueding ta zhe-xiang gongzuo mingtian wancheng  
 Xiaoming decide 3SG this-CL task tomorrow  
 wancheng  
 finish  
 ‘Xiaoming decides that he will finish this task tomorrow.’

**C-structure**



**F-structure**

PRED	'decide<[7:Xiaoming], [1:finish]>'										
COMP	<table border="1"> <tr> <td>PRED</td> <td>'finish&lt;[6:PRO], [5:task]&gt;'</td> </tr> <tr> <td>DIS</td> <td>{ PRED 'task' DEIXIS proximal, DEF + }</td> </tr> <tr> <td>ADJUNCT</td> <td>{ PRED 'tomorrow' }</td> </tr> <tr> <td>OBJ</td> <td>[5]</td> </tr> <tr> <td>SUBJ</td> <td>PRED 'PRO'</td> </tr> </table>	PRED	'finish<[6:PRO], [5:task]>'	DIS	{ PRED 'task' DEIXIS proximal, DEF + }	ADJUNCT	{ PRED 'tomorrow' }	OBJ	[5]	SUBJ	PRED 'PRO'
PRED	'finish<[6:PRO], [5:task]>'										
DIS	{ PRED 'task' DEIXIS proximal, DEF + }										
ADJUNCT	{ PRED 'tomorrow' }										
OBJ	[5]										
SUBJ	PRED 'PRO'										
SUBJ	PRED 'Xiaoming'										

**Test case 8: Partial subject control with the inner topic crossing control verb (expressed embedded subject)**

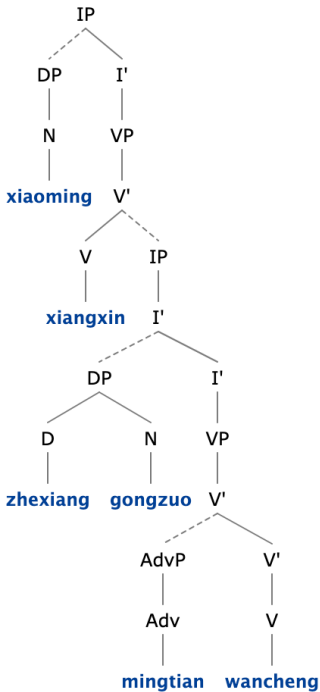
- (8) xiaoming zhe-xiang gongzuo ta jueding mingtian  
 Xiaoming this-CL task 3SG decide tomorrow  
 wancheng  
 finish  
 'Xiaoming decides that he will finish this task tomorrow.'

No formal solution could be produced.

**Test case 9: Non-control verb with the inner topic residing in the complement clause**

- (9) xiaoming xiangxin zhe-xiang gongzuo mingtian wancheng  
 Xiaoming believe this-CL task tomorrow finish  
 ‘Xiaoming believes that (he) will finish this task tomorrow.’

**C-structure**



**F-structure**

PRED	'believe<[7:Xiaoming], [1:finish]>'
PRED	'finish<[6:PRO], [5:task]>'
DIS	{ PRED 'task' DEIXIS proximal, DEF + }
COMP	ADJUNCT { 3   PRED 'tomorrow' }
OBJ	[5]
SUBJ	6   PRED 'PRO'
PS_LDD	-
SUBJ	7   PRED 'Xiaoming'

**Test case 10: Non-control verb with the inner topic crossing the non-control verb**

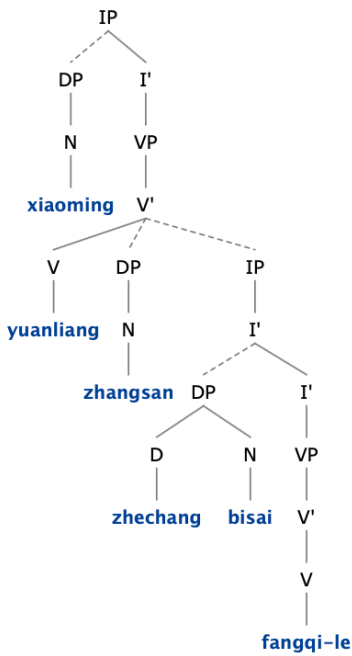
- (10) xiaoming zhe-xiang gongzuo xiangxin mingtian wancheng  
 Xiaoming this-CL task believe tomorrow finish  
 ‘Xiaoming believes that (he) will finish this task tomorrow.’

No formal solution could be produced.

Test case 11: Object-control verb with the inner topic residing in the complement clause

- (11) xiaoming yuanliang zhangsan zhe-chang bisai  
 Xiaoming forgive Zhangsan this-CL competition  
 fangqi-le  
 give.up-PFV  
 ‘Xiaoming forgives Zhangsan for giving up this competition.’

**C-structure**



**F-structure**

<b>PRED</b>	'forgive<[2:Xiaoming], [1:Zhangsan], [3:give-up]>'										
<b>XCOMP</b>	<table border="1"> <tr> <td><b>PRED</b></td> <td>'give-up&lt;[1:Zhangsan], [5:competition]&gt;'</td> </tr> <tr> <td><b>DIS</b></td> <td>{  <table border="1"> <tr> <td><b>PRED</b></td> <td>'competition'</td> </tr> <tr> <td><b>DEIXIS</b></td> <td>proximal,</td> </tr> <tr> <td><b>DEF</b></td> <td>+</td> </tr> </table> </td> </tr> </table>	<b>PRED</b>	'give-up<[1:Zhangsan], [5:competition]>'	<b>DIS</b>	{ <table border="1"> <tr> <td><b>PRED</b></td> <td>'competition'</td> </tr> <tr> <td><b>DEIXIS</b></td> <td>proximal,</td> </tr> <tr> <td><b>DEF</b></td> <td>+</td> </tr> </table>	<b>PRED</b>	'competition'	<b>DEIXIS</b>	proximal,	<b>DEF</b>	+
<b>PRED</b>	'give-up<[1:Zhangsan], [5:competition]>'										
<b>DIS</b>	{ <table border="1"> <tr> <td><b>PRED</b></td> <td>'competition'</td> </tr> <tr> <td><b>DEIXIS</b></td> <td>proximal,</td> </tr> <tr> <td><b>DEF</b></td> <td>+</td> </tr> </table>	<b>PRED</b>	'competition'	<b>DEIXIS</b>	proximal,	<b>DEF</b>	+				
<b>PRED</b>	'competition'										
<b>DEIXIS</b>	proximal,										
<b>DEF</b>	+										
	4 5										
<b>OBJ</b>	[5]										
<b>SUBJ</b>	1   PRED 'Zhangsan'										
	3   PS_LDD -, ASPECT perfective										
<b>OBJ</b>	[1]										
<b>SUBJ</b>	2   PRED 'Xiaoming'										
	0										

**Test case 12: Object-control verb with the inner topic crossing the object-control verb**

- (12) xiaoming zhe-chang bisai yuanliang zhangsan  
Xiaoming this-CL competition forgive Zhangsan  
fangqi-le  
give.up-PFV  
'Xiaoming forgives Zhangsan for giving up this competition.'

No formal solution could be produced to characterise *zhe-chang bisai* 'this competition' as the displaced object of *fangqi-le* 'give.up-PFV'.

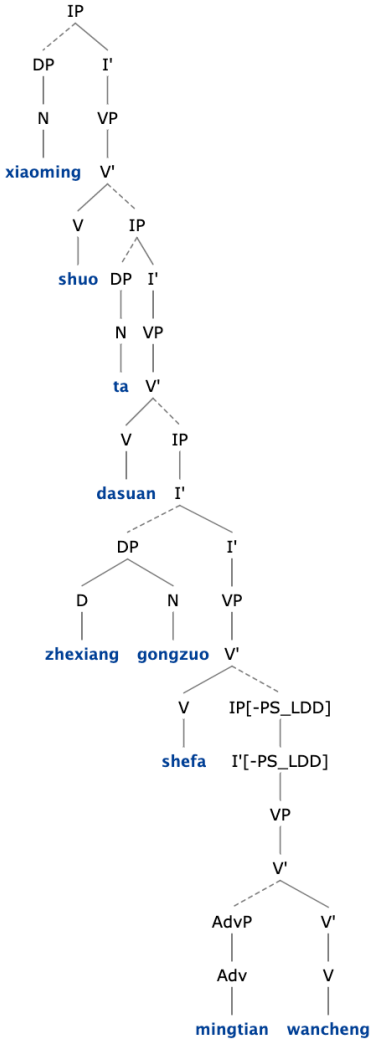
**Test cases 13–16: Complex embedding**

- (13) xiaoming shuo ta dasuan shefa zhe-xiang gongzuo  
Xiaoming say 3SG intend try this-CL task  
mingtian wancheng  
tomorrow finish  
'Xiaoming says he intends to try to finish this task tomorrow.'

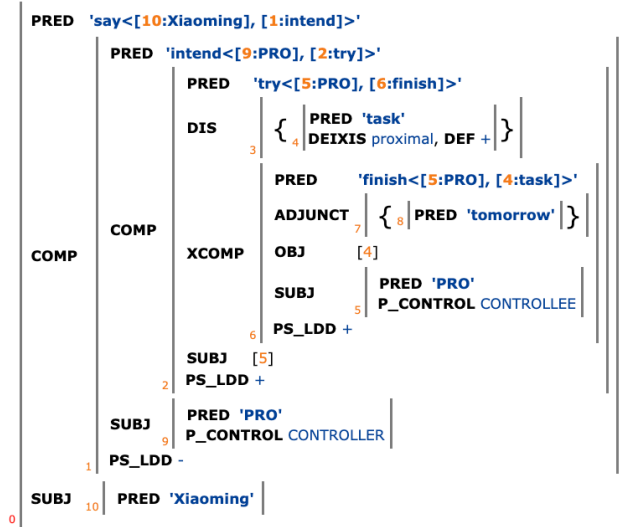
No formal solution could be produced.

- (14) xiaoming shuo ta dasuan zhe-xiang gongzuo shefa  
Xiaoming say 3SG intend this-CL task try  
mingtian wancheng  
tomorrow finish  
'Xiaoming says he intends to try to finish this task tomorrow.'

**C-structure**



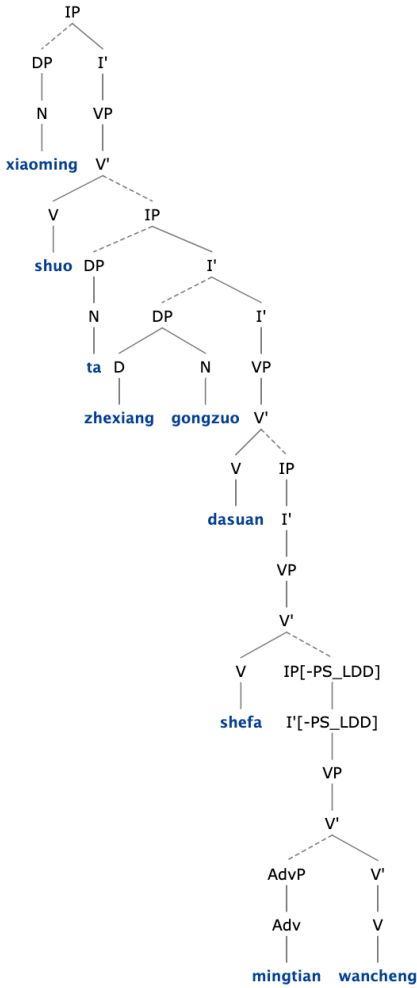
**F-structure**



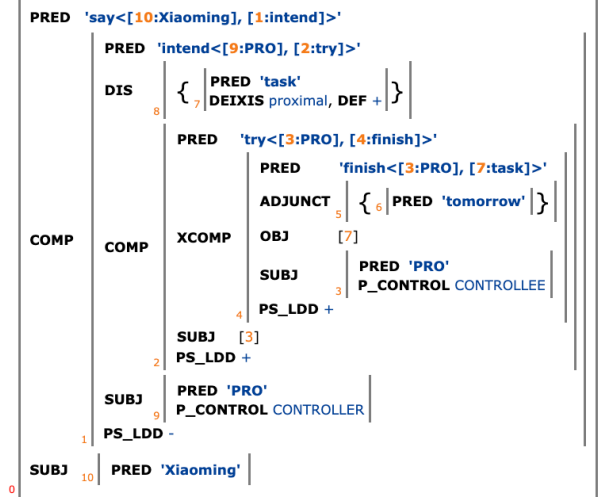
- (15) xiaoming shuo ta zhe-xiang gongzuo dasuan shefa  
 Xiaoming say 3SG this-CL task intend try  
 mingtian wancheng  
 tomorrow finish  
 'Xiaoming says he intends to try to finish this task tomorrow.'



**C-structure**



**F-structure**



- (16) xiaoming zhe-xiang gongzuo shuo ta dasuan shefa  
 Xiaoming this-CL task say 3SG intend try  
 mingtian wancheng  
 tomorrow finish  
 ‘Xiaoming says he intends to try to finish this task tomorrow.’

No formal solution could be produced.

## REFERENCES

- Judith AISSSEN and David PERLMUTTER (1976), Clause reduction in Spanish, in Henry THOMPSON, Kenneth WHISTLER, Vicki EDGE, Jeri J. JAEGER, Ronya JAVKIN, Miriam PETRUCK, Christopher SMEALL, and Robert D. VAN VALIN JR, editors, *Proceedings of the Second Annual Meeting of the BLS*, pp. 1–30.
- Ben AMBRIDGE and Adele E. GOLDBERG (2008), The island status of clausal complements: Evidence in favor of an information structure explanation, *Cognitive Linguistics*, 19(3), doi:10.1515/COGL.2008.014.
- Avery D. ANDREWS (1982), The representation of case in Modern Icelandic, in Joan BRESNAN, editor, *The mental representation of grammatical relations*, MIT Press.
- Ash ASUDEH (2005), Control and semantic resource sensitivity, *Journal of Linguistics*, 41(3):465–511, doi:10.1017/S0022226705003427.
- Ash ASUDEH (2009), Adjacency and locality: A constraint-based analysis for complementizer-adjacent extraction, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG09 Conference*.
- Markus BADER and Jana HÄUSSLER (2010), Toward a model of grammaticality judgments, *Journal of Linguistics*, 46(2):273–330, doi:10.1017/S0022226709990260.
- Emily BENDER (2008), Grammar engineering for linguistic hypothesis testing, in Nicholas GAYLORD, Alexis PALMER, and Elias PONVERT, editors, *Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less-Studied Languages*, pp. 16–36, CSLI Publications.
- Timothée BERNARD and Grégoire WINTERSTEIN (2022), Introduction to the special section on the interaction between formal and computational linguistics, *Journal of Language Modelling*, 10(1):39–47, doi:10.15398/jlm.v10i1.325.
- Joan BRESNAN (1982), Control and complementation, *Linguistic Inquiry*, 13(3):343–434.
- Joan BRESNAN, Ash ASUDEH, Ida TOIVONEN, and Stephen WECHSLER (2016), *Lexical-Functional Syntax*, John Wiley & Sons.
- Fabian BROSS (2019), Using mixed effects models to analyze acceptability rating data in linguistics, <https://www.fabianbross.de/mixedmodels.pdf>.
- Miriam BUTT (2014), Control vs. complex predication: Identifying non-finite complements, *Natural Language & Linguistic Theory*, 32(1):165–190, doi:10.1007/s11049-013-9217-5.
- Miriam BUTT, Tracy Holloway KING, María-Eugenia NIÑO, and Frédérique SEGOND (1999), *A grammar writer's cookbook*, CSLI Publications.

- Kersti BÖRJARS, Christopher HICKS, and John PAYNE (2018), Interdependencies in Chinese noun phrases, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of the LFG'18 Conference*, pp. 109–128, CSLI Publications.
- Kersti BÖRJARS, Rachel NORDLINGER, and Louisa SADLER (2019), *Lexical-Functional Grammar: An introduction*, Cambridge University Press, doi:10.1017/9781316756584.
- Hilary CHAPPELL (2008), Variation in the grammaticalization of complementizers from *verba dicendi* in Sinitic languages, *Linguistic Typology*, 12(1), doi:10.1515/LITY.2008.032.
- Isabelle CHARNAVEL, C.-T. James HUANG, Peter COLE, and Gabriella HERMON (2017), Long-distance anaphora: Syntax and discourse, in *The Wiley Blackwell Companion to Syntax*, pp. 1–82, John Wiley & Sons, Inc.
- Dewei CHE and Adams BODOMO (2018), A Constraint-based analysis of the objects of VO compounds in Mandarin Chinese, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of LFG18 Conference*, pp. 172–190, CSLI Publications.
- Rune H. B. CHRISTENSEN (2020), Ordinal – Regression models for ordinal data, <https://CRAN.R-project.org/package=ordinal>.
- Guglielmo CINQUE (2006), *Restructuring and functional heads*, Oxford University Press.
- Dick CROUCH, Mary DALRYMPLE, Ronald M. KAPLAN, Tracy Holloway KING, John MAXWELL, and Paula NEWMAN (2011), *XLE documentation*, Palo Alto Research Centre.
- Mary DALRYMPLE, John J. LOWE, and Louise MYCOCK (2019), *The Oxford reference guide to Lexical Functional Grammar*, Oxford University Press.
- Denys DUCHIER and Yannick PARMENTIER (2015), High-level methodologies for grammar engineering, introduction to the special issue, *Journal of Language Modelling*, 3(1):5–19, doi:10.15398/jlm.v3i1.117.
- Thomas ERNST and Chengchi WANG (1995), Object preposing in Mandarin Chinese, *Journal of East Asian Linguistics*, 4(3):235–260.
- Nomi ERTESCHIK (1973), *On the nature of island constraints*, PhD Thesis, Massachusetts Institute of Technology.
- Martin FORST and Tracy Holloway KING (2023), Computational implementations and applications, in Mary DALRYMPLE, editor, *The handbook of Lexical Functional Grammar*, Language Science Press.
- John FOX and Sanford WEISBERG (2019), *An R companion to applied regression*, Sage, 3rd edition, <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

- Grant GOODALL, editor (2021a), *The Cambridge handbook of experimental syntax*, Cambridge University Press, doi:10.1017/9781108569620.
- Grant GOODALL (2021b), Sentence acceptability experiments: What, how, and why, in Grant GOODALL, editor, *The Cambridge handbook of experimental syntax*, pp. 7–38, Cambridge University Press.
- Thomas GRANO (2015), *Control and restructuring*, Oxford University Press.
- Thomas GRANO and Howard LASNIK (2018), How to neutralize a finite clause boundary: Phase theory and the grammar of bound pronouns, *Linguistic Inquiry*, 49(3):465–499, doi:10.1162/ling\_a\_00279.
- Dag HAUG (2013), Partial control and anaphoric control in LFG, in Miriam BUTT and Tracy Holloway KING, editors, *Proceedings of LFG13*, pp. 274–294, CSLI Publications.
- Dag HAUG (2014), The anaphoric semantics of partial control, in Todd SNIDER, Sarah D’ANTONIO, and Mia WEIGAND, editors, *Proceedings of SALT*, pp. 213–233.
- One-Soon HER (2009), Unifying the long passive and the short passive: On the bei construction in Taiwan Mandarin, *Language and Linguistics*, 10(3):421–470.
- One-Soon HER (2012), Structure of classifiers and measure words: A lexical functional account, *Language and Linguistics*, 13(6):1211–1251.
- Maxime R. HERVÉ (2022), RVAideMemoire: Testing and plotting procedures for biostatistics, <https://CRAN.R-project.org/package=RVAideMemoire>.
- Jianhua HU, Haihua PAN, and Liejiong XU (2001), Is there a finite vs. nonfinite distinction in Chinese?, *Linguistics*, 39(6):1117–1148.
- C.-T. James HUANG (1984), On the distribution and reference of empty pronouns, *Linguistic Inquiry*, 15(4):531–574.
- C.-T. James HUANG (1989), Pro-drop in Chinese: A generalized control theory, in Osvaldo A. JAEGGLI and Kenneth J. SAFIR, editors, *The null subject parameter*, pp. 185–214, Springer.
- C.-T. James HUANG, Yen-Hui Audrey LI, and Yafei LI (2009), *The syntax of Chinese*, Cambridge University Press.
- Nick HUANG (2018), Control complements in Mandarin Chinese: implications for restructuring and the Chinese finiteness debate, *Journal of East Asian Linguistics*, 27(4):347–376.
- Nick HUANG (2021), How subjects and possessors can obviate phasehood, *Linguistic Inquiry*, pp. 1–32, doi:10.1162/ling\_a\_00414.
- Nick HUANG, Diogo ALMEIDA, and Jon SPROUSE (2022), How good are leading theories of bridge verbs? an experimental evaluation, [https://z-n-huang.github.io/files/presentations/HuangAlmeidaSprouse2022\\_WCCFL\\_bridge\\_verbs.pdf](https://z-n-huang.github.io/files/presentations/HuangAlmeidaSprouse2022_WCCFL_bridge_verbs.pdf).

- Ray JACKENDOFF (1977), *X-Bar syntax: A study of phrase structure*, MIT Press.
- Miloš JAKUBÍČEK, Adam KILGARRIFF, Vojtěch KOVÁŘ, Pavel RYCHLÝ, and Vít SUCHOMEL (2013), The TenTen corpus family, pp. 125–127, Lancaster University.
- Ronald M. KAPLAN and Joan BRESNAN (1982), Lexical-Functional Grammar: A formal system for grammatical representation, in Joan BRESNAN, editor, *The mental representation of grammatical relations*, pp. 173–281, MIT Press.
- Ronald M. KAPLAN and Annie ZAENEN (1989), Long-distance dependencies, constituent structure, and functional uncertainty, in Mark BALVIN and Anthony KROCH, editors, *Alternative Conceptions of Phrase Structure*, pp. 17–42.
- Chit-Fung LAM (2021), A constraint-based approach to anaphoric and logophoric binding in Mandarin Chinese and Cantonese, in Miriam BUTT, Jamie FINDLAY, and Ida TOIVONEN, editors, *Proceedings of the LFG'21 Conference*, pp. 202–222, CSLI Publications.
- Chit-Fung LAM (2022), Rethinking restructuring in Mandarin Chinese: Empirical properties, theoretical insights, and LFG/XLE computational implementation, in Miriam BUTT, Jamie FINDLAY, and Ida TOIVONEN, editors, *Proceedings of the LFG'22 Conference*, pp. 203–222, CSLI Publications.
- Chit-Fung LAM (2023), *Control and complementation in parallel constraint-based architecture: An empirically oriented investigation of Mandarin Chinese*, PhD Thesis, University of Manchester.
- Idan LANDAU (2000), *Elements of control: Structure and meaning in infinitival constructions*, Kluwer Academic Publishers.
- Idan LANDAU (2013), *Control in generative grammar: A research companion*, Cambridge University Press.
- Charles N. LI and Sandra A. THOMPSON (1989), *Mandarin Chinese: A functional reference grammar*, University of California Press.
- Yingtong LIU, Rachel RYSKIN, Richard FUTRELL, and Edward GIBSON (2022), A verb-frame frequency account of constraints on long-distance dependencies in English, *Cognition*, 222:104902, doi:10.1016/j.cognition.2021.104902.
- John LOWE and Joseph LOVSTRAND (2020), Minimal phrase structure: a new formalized theory of phrase structure, *Journal of Language Modelling*, 8(1):1–352, doi:https://doi.org/10.15398/jlm.v8i1.247.
- Karuvannur P. MOHANAN (1983), Functional and anaphoric control, *Linguistic Inquiry*, 14(4):641–674.
- Stefan MÜLLER (2015), The CoreGram project: theoretical linguistics, theory development and verification, *Journal of Language Modelling*, 3(1):21–86, doi:10.15398/jlm.v3i1.91.
- Marie-Claude PARIS (1998), Focus operators and types of predication in Mandarin, *Cahiers de Linguistique – Asie Orientale*, 27(2):139–159.

- Waltraud PAUL (2002), Sentence-internal topics in Mandarin Chinese: The case of object preposing, *Language and Linguistics*, 3(4):695–714.
- Waltraud PAUL (2005), Low IP area and left periphery in Mandarin Chinese, *Recherches Linguistiques de Vincennes*, 33:111–133.
- Waltraud PAUL (2015), *New perspectives on Chinese syntax*, De Gruyter.
- Yanfeng QU (1995), *Object noun phrase dislocation in Mandarin Chinese*, PhD Thesis, University of British Columbia.
- Luigi RIZZI (1978), A restructuring rule in Italian syntax, in Samuel KEYSER, editor, *Recent transformational studies in European languages*, pp. 113–158, MIT Press.
- Victoria ROSÉN, Koenraad DE SMEDT, Paul MEURER, and Helge DYVIK (2012), An open infrastructure for advanced treebanking, in Jan HAJIČ, Koenraad DE SMEDT, Marko TADIĆ, and António BRANCO, editors, *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pp. 22–29, LREC2012 Workshop.
- Shu-ing SHYU (1995), *The syntax of focus and topic in Mandarin Chinese*, PhD Thesis, University of Southern California Los Angeles.
- Michaela SPITZER, Jan WILDENHAIN, Juri RAPPILBER, and Mike TYERS (2014), BoxPlotR: A web tool for generation of box plots, *Nature Methods*, 11(2):121–122, doi:10.1038/nmeth.2811.
- Jon SPROUSE and Diogo ALMEIDA (2012), Power in acceptability judgment experiments and the reliability of data in syntax, <https://lingbuzz.net/lingbuzz/001520>.
- Barbara STIEBELS (2007), Towards a typology of complement control, in Barbara STIEBELS, editor, *Studies in complement control. ZAS papers in Linguistics*, pp. 1–80.
- Sebastian SULGER, Miriam BUTT, Tracy Holloway KING, Paul MEURER, Tibor LAZCKO, Gyorgy RAKOSI, Cheikh Bamba DIONE, Helge DYVIK, Victoria ROSÉN, Koenraad DE SMEDT, Agnieszka PATEJUK, Ozlem CETINOGLU, Wayan I ARKA, and Meladel MISTICA (2013), Pargrambank: The pargram parallel treebank, in Hinrich SCHUETZE, Pascale FUNG, and Massimo POESIO, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 550–560, OmniPress.
- Aaron Steven WHITE and Thomas GRANO (2014), An experimental investigation of partial control, in Urtzi ETXEBERRIA, Anamaria FĂLĂUȘ, Aritz IRURTZUN, and Bryan LEFERMAN, editors, *Proceedings of Sinn Und Bedeutung*, pp. 469–486.
- Susi WURMBRAND (2001), *Infinitives: Restructuring and clause structure*, De Gruyter, doi:10.1515/9783110908329.

*Control, inner topicalisation, and focus fronting*

Susi WURMBRAND (2004), Two types of restructuring – Lexical vs. functional, *Lingua*, 114(8):991–1014, doi:10.1016/S0024-3841(03)00102-5.

Susi WURMBRAND (2015), Restructuring cross-linguistically, in Thuy BUI and Deniz ÖZYILDIZ, editors, *Proceedings of NELS 45*, University of Massachusetts.

Susi WURMBRAND and Magdalena LOHNINGER (2019), An implicational universal in complementation: Theoretical insights and empirical progress, in Jutta M. HARTMANN and Angelika WÖLLSTEIN, editors, *Propositionale Argumente im Sprachvergleich: Theorie und Empirie [Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues]*, Gunter Narr Verlag, Tübingen.

Annie ZAENEN (1980), *Extraction rules in Icelandic*, PhD Thesis, Harvard University.

Olga ZAMARAEVA, Chris CURTIS, Guy EMERSON, Antske FOKKENS, Michael GOODMAN, Kristen HOWELL, T.J. TRIMBLE, and Emily M. BENDER (2022), 20 years of the Grammar Matrix: cross-linguistic hypothesis testing of increasingly complex interactions, *Journal of Language Modelling*, 10(1):49–137, doi:10.15398/jlm.v10i1.292.

Niina Ning ZHANG (2016), Identifying Chinese dependent clauses in the forms of subjects, *Journal of East Asian Linguistics*, 25(3):275–311, doi:10.1007/s10831-016-9146-5.

*Chit-Fung Lam*

© 0000-0001-5094-2627

chitfung.lam@manchester.ac.uk

Linguistics and English Language,  
School of Arts, Languages and Cultures,  
University of Manchester,  
Oxford Road, Manchester, M13 9PL, UK

Chit-Fung Lam (2024), *Control, inner topicalisation, and focus fronting in Mandarin Chinese: modelling in parallel constraint-based grammatical architecture*, *Journal of Language Modelling*, 12(1):69–153

doi <https://dx.doi.org/10.15398/jlm.v12i1.365>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

© <http://creativecommons.org/licenses/by/4.0/>





# On German verb sense disambiguation: A three-part approach based on linking a sense inventory (GermaNet) to a corpus through annotation (TGVCorp) and using the corpus to train a VSD classifier (TTvSense)

*Dominik Mattern*<sup>1</sup>, *Wahed Hemati*<sup>2</sup>, *Andy Lücking*<sup>1</sup>, and *Alexander Mehler*<sup>1</sup>

<sup>1</sup> Goethe University Frankfurt, Text Technology Lab

<sup>2</sup> Shikenso GmbH

## ABSTRACT

We develop a three-part approach to Verb Sense Disambiguation (VSD) in German. After considering a set of lexical resources and corpora, we arrive at a statistically motivated selection of a subset of verbs and their senses from GermaNet. This sub-inventory is then used to disambiguate the occurrences of the corresponding verbs in a corpus resulting from the union of TüBa-D/Z, Sa1sa, and E-VALBU. The corpus annotated in this way is called TGVCorp. It is used in the third part of the paper for training a classifier for VSD and for its comparative evaluation with a state-of-the-art approach in this research area, namely EWISER. Our simple classifier outperforms the transformer-based approach on the same data in both accuracy and speed in German but not in English and we discuss possible reasons.

*Keywords:*  
*verb sense*  
*disambiguation*  
*(VSD),*  
*word sense*  
*disambiguation*  
*(WSD)*

Ambiguity arises when a word or a multi-word constituent is associated with more than one meaning (Chierchia and McConnell-Ginet 2000, p. 38; see Kennedy 2011 for an overview). The multiple meanings of a word are referred to as *senses*. Choosing just one from the many senses of an ambiguous word in context is a process known as Word Sense Disambiguation (WSD) (Navigli 2009). Here we focus on *Verb Sense Disambiguation* (VSD), i.e., selecting a sense from the sense enumerations associated with a given verb. We present an approach to the disambiguation of German verbs. We briefly set the theoretical stage in Section 1.1 and review related NLP work in Section 1.2.

## 1.1

*Ambiguity and context variability*

VSD is a lexical issue: determining which of the verb's senses is appropriate in a given context.<sup>1</sup> Lexical ambiguity is expressed in terms of word sense enumerations: each meaning of an ambiguous word corresponds to one sense. Traditionally, lexical ambiguity is attributed to either polysemy (a single word form is associated with various senses) or homonymy (different senses happen to share the same orthographic (homograph) or phonological (homophone) representation) (Lyons 1977, p. 550). The two varieties of lexical ambiguity can be difficult to distinguish (though there are some guidelines, see Kroeger 2019, Section 5.3.3). Verb ambiguity is illustrated in (1), taken from Cruse (2000, p. 108):

- (1) a. John expired last Thursday.  
 b. John's driving licence expired last Thursday.  
 c. ?John and his driving licence expired last Thursday.

---

<sup>1</sup>Thus, verbs exhibit *lexical ambiguity*. Other types of ambiguity known from nouns and adjectives and the phrases constructed out of those parts of speech are syntactic or structural ambiguity (*competent men and women*; Chierchia and McConnell-Ginet 2000, p. 38), as well as scope ambiguity (*Every schoolgirl crossed a road*; Dwivedi 2013).

The proper name *John* in (1a) calls for an interpretation of the verb *expire* in terms of “dying”, while in (1b) an “end of period” reading is selected. Linguistic evidence for the polysemy of *expiring* is exemplified in (1c) (the question mark indicates semantic oddity): In the *antagonism test* (Kroeger 2019, Section 5.3.2), only different senses lead to the zeugma effect (the effect that the verb senses of conjoined verbs are antagonistic; for ambiguity tests see Zwicky and Sadock 1975; see Gillon 1990 for some critical discussion).

Disambiguation relies heavily on context information. For instance, keeping the two senses of *expiring* apart in (1) is based on world knowledge about proper names of persons and bureaucratic administrations. Accordingly, it is important to distinguish ambiguity from the general context variability of meanings (Cruse 2000, Chapter 6).<sup>2</sup> Let us illustrate the subtle differences between polysemy and context-variability by means of a positive and a negative example each. Consider the following sentences from German (since we are concerned with German VSD):

- (2) a. Das Gerät **läuft** einwandfrei. (*The device works correctly.*)  
b. Der Schaffner **läuft** zum Bahnhof. (*The ticket collector walks to the station.*)  
c. ?Das Gerät **läuft** und der Schaffner auch. (? *The device is running and so is the ticket collector.*)

The verb form *läuft* has two different meanings in sentences (2a) and (2b), which can be paraphrased with “it works” and “it walks”, respectively. It is noteworthy, but by no means a rule, that the same German word form receives a different English translation for each sense. For this reason, we will have a particular focus on multilingual

---

<sup>2</sup> Context-sensitive effects of contents include indexicality (the first person pronoun *I*, for example, is not ambiguous despite referring to a potentially different person on each occasion of use; Kaplan 1989), coercion (e.g., type-shifting the noun *novel* to an eventive argument in *He began the novel*; Moens and Steedman 1988; Pustejovsky 1995; de Swart 2011), co-composition or co-predication (as observed, for instance, with “interactive verb-argument compositions” such as *Pat swallowed the lemonade* vs. *Pat swallowed her worries*; Pustejovsky 1991, 1995; Asher et al. 2017; Cooper 2011).

WSD resources. (2c) shows that polysemy is indicated by the antagonism test, which leads to a zeugma effect. The two senses are correctly kept apart in our approach.

However, *laufen* ‘to run’ can also be used to denote directed or undirected movement (Jackendoff 1983):

- (3) a. Er *läuft* so schnell es geht zum Zug. (*He runs to the train as fast as possible; run<sub>1</sub> = go-to(x,y)*)  
b. Sie *läuft* durch den Park. (*She runs through the park; run<sub>2</sub> = move(x)*)  
c. Sie *laufen* zum Zug und durch den Park. (*They run to the train and through the park.*)

In contrast to (2), *laufen* ‘to run’ in (3) passes the antagonism test without giving rise to a zeugma effect, which provides evidence for a shared verb sense in both conjuncts. Furthermore, both verb occurrences are translated to the same English word form. With regard to semantics, both directed and undirected movements follow from interactive meaning composition (Pustejovsky 1991), so no sense enumeration is needed. Thus the pattern in (3) is due to a single sense of the verb. Since (3a) and (3b) are attributed to different senses in our account, we observe some overgeneralization of lexical ambiguity.

What about figurative language use such as metaphor or metonymy? Cruse (2000, p. 112) puts them among polysemy, namely as non-linear types of polysemy.<sup>3</sup> However, this classification lacks empirical support: metonymic uses of a noun phrase, for instance, do not seem to rest on ambiguity, but rather on a “transfer of meaning” (predicate transfer, in this case) (Nunberg 1995).<sup>4</sup> Consequently, we take figurative speech to be a matter of inference, not of WSD.

A note on terminology: We use the terms “valence” or “subcategorization” for the syntactic arguments of a verb. For example, a tran-

---

<sup>3</sup>They are non-linear because they lack a linear specialization relationship towards their “siblings”.

<sup>4</sup>To briefly rehash one of Nunberg’s arguments: the noun phrase *ham sandwich*, even when used metonymically in a restaurant in order to refer to its orderer, still preserves its basic meaning since it can be picked out by discourse anaphora: *The ham sandwich seems to be enjoying it* (*it = the ham sandwich*).

sitive verb such as *eat* takes a subject and a complement – hence, there are two noun phrases on its valence or subcategorization list. These elements are mapped onto the verb’s argument structure and linked to content representations (linking) (Wechsler *et al.* 2021). There are different approaches to representing contents; we will refer to semantic arguments of content representations as *semantic roles*.<sup>5</sup>

### VSD for German

1.2

Word Sense Disambiguation (WSD) in general is essential for many (if not all) Natural Language Processing (NLP) applications that require semantic information. The disambiguation of verbs, VSD, is of particular importance when it comes to Semantic Role Labeling (SRL) (Palmer *et al.* 2010). This is due to the fact that the argument structure or subcategorization frame of verbs can differ with their senses. Consider again *laufen* ‘to run’ from (2). While (2a) and (2b) select for a nominal nominative subject, the subject is linked differently to the semantic arguments provided by the verb sense-specific predication. Such argument structure linking can be achieved in various ways including selectional restrictions (e.g.  $\pm$ ANIMATE) (Soehn 2005) or lexical frames (respectively parameterized states of affairs; e.g. *operating-frame* vs. *movement-frame*) (Wechsler *et al.* 2021).<sup>6</sup> Thus, if the representation of meaning fails already on the level of verb occurrences in sentences, because it is not able to distinguish between different senses connected with the same form, then a precondition for determining the corresponding sentence meaning is missing (Levin 1993). This leads us to the assessment that any reasonable approach to sentence or text meaning representation (which goes beyond black box

---

<sup>5</sup>WSD approaches usually refrain from using argument structures in the grammar-theoretic sense and employ a direct mapping from syntactic arguments to semantic representations, as is done in Semantic Role Labeling (SRL). Hence, the term “argument structure” when used in these contexts is to be understood either in terms of syntactic subcategorization or semantic roles.

<sup>6</sup>Resources used for SRL differ in the granularity and nomenclature of their argument vocabularies. A recent resource addresses this inter-operability issue by providing yet another synset-based vocabulary but with links to FrameNet (Fillmore and Baker 2010), VerbNet (Schuler 2006), PropBank (Bonial *et al.* 2015) and WordNet (Fellbaum and Miller 1998) roles (Di Fabio *et al.* 2019).

models based e.g. on current neural networks) must perform VSD as a preprocessing step. Hence, there is already a history of lexical representations and WSD, including lexical resources (Miller 1995; Schuler 2006; Baker *et al.* 1998) and sense annotated corpora (Edmonds and Cotton 2001; Snyder and Palmer 2004; Pradhan *et al.* 2007; Navigli *et al.* 2013).

However, existing resources focus on English; there is little research on WSD in high resource languages such as German, especially for verbs. German WSD was featured on SemEval as a task or partial task only twice (Lefever and Hoste 2010, 2013), in both cases as part of a multilingual disambiguation task only involving a small number of nouns (see Figure 1).

To promote NLP for or based on SRL and related tasks in German, a correspondingly large dataset with high verb lemma coverage and a standardized sense inventory is needed. The present work aims to fill this gap by means of a three-layer architecture of VSD which integrates (1) the modeling and post-processing of verb sense representations with (2) the generation of training data annotation and (3) the machine learning based thereon. This approach, first elaborated in Hemati (2020) and considerably extended and further validated here, is compared in detail with related resources below. Such resources have been provided in few previous works on German verbs (for an evaluation of WSD algorithms for German *nouns* see Henrich and Hinrichs 2012):

1. The “Elektronische Valenzwörterbuch” (*electronic valence dictionary*) of German verbs, E-VALBU (Kubczak 2009), contains the 638 verbs from the printed VALBU (Schumacher *et al.* 2004), plus 30 new verb lemmas from the domain of a general science vocabulary. Grammatical descriptions and disambiguation of the E-VALBU verbs are based on their usage context in DEREKO (Dipper *et al.* 2002) and are obtained using corpus-assisted lexicographical methods (Schumacher 1986). For that reason, E-VALBU, though being a reference corpus, is of limited coverage.
2. Scheible *et al.* (2013) developed a rule-based *SubCat-Extractor*, which obtains subcategorization information from parsed corpora annotated with STTS (Schiller *et al.* 1999) such as the TIGER corpus (Brants *et al.* 2004). The SubCat-Extractor was applied

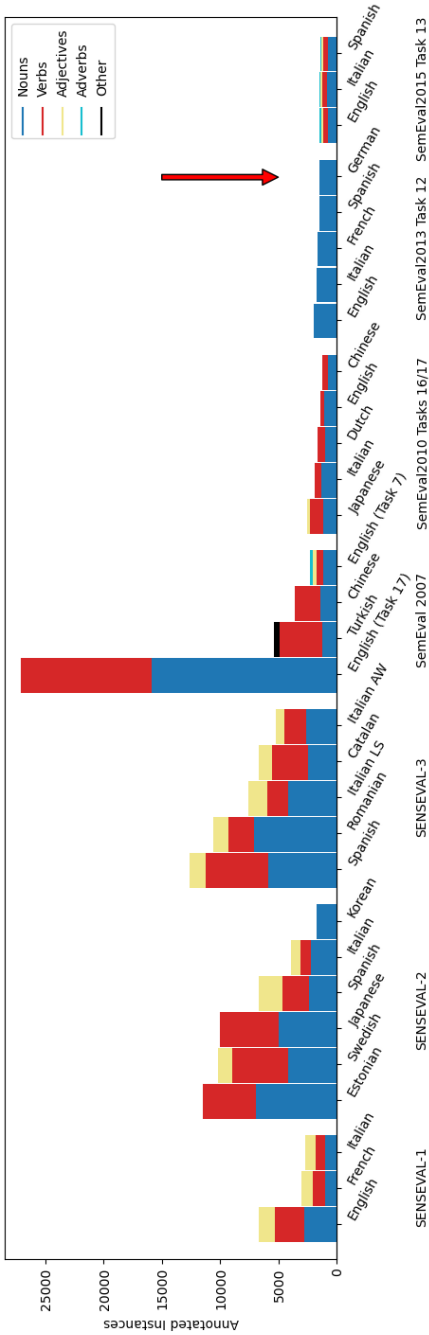


Figure 1: Number of annotated instances in the training and test corpora used for WSD during the SENSEVAL/SemEval Tasks. The SENSEVAL/SemEval tasks are a series of evaluations designed to assess the effectiveness of computational systems in understanding the meaning (or “sense”) of words in different contexts, crucial for word sense disambiguation (WSD). Only supervised WSD tasks are included. Some of the languages are missing from SENSEVAL-2 and -3, specifically Basque, Czech, Dutch and English from SENSEVAL-2 and Basque, Chinese and English tasks from SENSEVAL-3. The papers for these languages did not include complete POS breakdowns of the corpora. In short, the distribution shows that verb-related WSD is a rare topic, especially for German

to SdeWac (Faaß and Eckart 2013). Although not explicitly connected to VSD, the resulting subcategorization lexicon of German verbs may contain different syntactic argument frames for a given verb, which often correlates with different semantic construals (as with the Levin 1993 classes). Since the verbs are retrieved from a large web-crawled database, the SubCat-Extractor resource has reasonable coverage. However, no explicit link to meaning labels is established.

3. VSD on a restricted class of verbs, namely perception verbs, was carried out by David *et al.* (2014). The focus of this paper was on distinguishing between perception verbs exhibiting literal and non-literal meanings. To this end, the authors selected one example of an optical, an acoustic, an olfactory, and a haptic verb each. The four verbs were assigned to 3 to 4 senses (1 literal and 2 to 3 non-literal), based on a corpus survey. Then a database was created by manually annotating 50 randomly chosen sentences for each selected perception verb in terms of the previously defined senses (i.e., 200 sentences in total). A decision tree was trained on the resulting dataset exploiting various features, partly drawing on the resource of Scheible *et al.* (2013). The classifier reached accuracies between 45.5 % and 69.4 %, however, due to the rather special focus of the approach it is difficult to generalize it to other VSD phenomena.
4. Henrich (2015) presents the most comprehensive work on VSD in German. She analyzed various corpora, including manually annotated and automatically created ones. In particular, she created a new German resource for WSD, namely WebCAGe (*Web-Harvested Corpus Annotated with GermaNet Senses*). WebCAGe rests on a semi-automatic alignment of Wiktionary glosses and GermaNet senses. Wiktionary was used to enlarge the set of sample sentences, most notably by exploiting links to Wikipedia articles. Following the “one sense per discourse” heuristics (Gale *et al.* 1992), occurrences of target words in external but linked sources are likely to be used in the same sense as that of the pivot word from a Wiktionary gloss. It should be noted that WebCAGe contains only words with more than one GermaNet sense, that is, words that are polysemous in GermaNet’s sense – unambiguous



words are excluded on purpose (since WebCAGE is designed as a *disambiguation* dataset). The resource creation process was semi-automatic, as the large-scale annotation is done automatically, followed by a manual post-correction. The resulting dataset was evaluated by lexicographers. The focus of WebCAGE, however, was on WSD (i.e. nouns, verbs, and adjectives). As a result, Henrich (2015) does not achieve high coverage for German verbs: the disambiguation resource includes 3,190 tagged verb tokens which belong to 897 polysemous verbs in GermaNet, exhibiting 3.6 verb senses on average (Henrich 2015, p. 118).<sup>7</sup>

5. A cross-lingual, multimodal approach to VSD was taken by Gella *et al.* (2019). They provide the MultiSense image dataset, which comprises 9,504 images annotated with English verbs and their translations into German and Spanish. MultiSense covers 55 English verbs with 154 (German) and 136 (Spanish) unique translations. The dataset is divided into 75% training, 10% validation and 15% test splits. The best performing model in a translation task was a mixed one which used visual and textual features. MultiSense departs from the sense enumeration paradigm (see Section 1.1) and delegates disambiguation to a translation process (namely translating the pivot verb into verbs of the remaining two target languages). Since the target language verbs are not disambiguated either, it is obvious that this approach only works for VSD if the target verbs are unambiguous – which is probably rarely the case (as a simple example reconsider (2)).<sup>8</sup>

In order to gain a better verb-related database for NLP in German beyond these resources, we created the TTLab German Verb Sense Corpus (TGVCORP). TGVCORP is a German corpus with a very high degree of coverage regarding the annotation of the senses of a high number of frequent verbs. Since the annotation of data is time-consuming and therefore cost-intensive, we developed a generic procedure to quickly

---

<sup>7</sup>In total WebCAGE contains 10,750 tagged word tokens which belong to 2,607 distinct polysemous words in GermaNet (Henrich 2015, p. 118).

<sup>8</sup>A further issue might reside in the *prima facie* appealing use of images as a *lingua franca*: While mundane, concrete actions can be depicted straightforwardly, it is difficult to see how more abstract contents such as those needed for attitude verbs are captured.

create high-quality training data for WSD. This procedure integrates three methods for the automatic generation of annotations employing translation models, language models and an inductive heuristics based on sense compression. TGVCorp contains manually annotated data for 1,560 ambiguous verb lemmas covering more than 78% of the verb tokens in COW (Schäfer and Bildhauer 2012), which is one of the largest openly accessible corpora for German. We use neural network-based tools for WSD and demonstrate their adaptation to VSD. We reproduce the experiments of Henrich (2015) and compare our approach with hers. In direct comparison to Henrich 2015, our most efficient model offers a performance increase of 8.4%, creating a new gold standard. We additionally present a simple method for generalizing senses that allows us to disambiguate verbs that are not present in the training set. With our approach, we achieve the highest verb token coverage for German VSD while maintaining state-of-the-art performance.

The paper is organized as follows: Section 2 describes TGVCorp and our procedure for creating it semi-automatically. Section 3 presents our supervised classifier for VSD based on TGVCorp. Finally, Section 4 concludes and discusses future work.

## 2 FROM RAW TEXTUAL DATA TO A SENSE-DISAMBIGUATED TEXT CORPUS: A THREE-LEVEL ARCHITECTURE

In this section we first describe the selection of the sense inventory underlying TGVCorp. We then turn to the generation of TGVCorp and evaluate its coverage using a larger set of different (genre- and topic-diverse) corpora. Finally, we describe the annotation of senses in this corpus, which are used in the remainder of the paper to train a supervised VSD classifier.

The significant expansion of annotation of verb senses in corpora is needed to train better classifiers for VSD. That is, instead of training new classifiers all the time, we rely on the idea of expanding the database and its quality to arrive at better NLP methods. To support

the generation of such a resource on the example of VSD, each target verb requires a list of its senses with sufficient information per sense so that they can be adequately captured, identified, and distinguished from each other by annotators. Creating our own list from scratch would be too complex, so we used existing inventories to gain a working basis. Hence, the first step was to determine which inventory is most appropriate for German VSD (Section 2.1). Likewise, we had to choose a corpus to start with, so in addition we examined several corpora (Section 2.2). Since human annotation is costly, we combined several methods to map the selected corpus to the selected inventory while minimizing annotation effort and keeping data quality high (Section 2.3).

### *Sense inventories*

2.1

A sense of a word  $w$  is a generally accepted meaning of  $w$  represented as a gloss, a paraphrase or as a synset in a WordNet (Fellbaum 1998). In a sense inventory these senses are enumerated per word. Independent of the question whether word senses can be enumerated as discretizable units, inventories map words to finite discrete sets of senses, each representing a certain meaning of the corresponding word. However, it is doubtful that there are periods of time in which the senses of a word can be completely discretized, so that one knows exactly where one sense begins and another ends (Rieger 1989, 2001). The discrete approach comes up against the fact that natural languages are permanently affected by change as a result of constantly changing contexts of language use (Keller 1990) – see Steels 2011–12 for a consideration of language dynamics from the point of view of evolutionary processes. This dynamic cannot be represented by sense lists, which are based on the implicit assumption of sufficiently stable senses, without actually measuring this stability: *Is the stability of the senses of words equally distributed? (Most likely not.) What does this stability depend on? Are the periods during which particular senses are observed sufficiently long so that a valid WSD can be performed? What does this mean for the selection of appropriate text corpora? Are these even sufficiently available for these periods?* Ideally, these and related questions should be clarified in order to make sense inventories a valid representation format.

Figure 2:  
Senses of the  
German verb  
*abtragen*  
'to dismantle'  
in two sense  
inventories:  
Duden  
(download:  
February 14,  
2024) (left)  
and Wiktionary  
(download:  
February 14,  
2024) (right)

Source: Duden	Source: Wiktionary
<i>abtragen / to dismantle</i>	<i>abtragen / to dismantle</i>
<b>Senses:</b> [1.a] Wiktionary[1] [1.b] Wiktionary[1] [1.c] Wiktionary[4] [2] Wiktionary[3] [3] Wiktionary[2] [4] Wiktionary[6]	<b>Senses:</b> [1] schichtweise entfernen [2] Kleidung so lange benutzen, bis sie kaputt ist [3] bezahlen [4] <i>Haushalt, gehoben</i> : das Geschirr vom Tisch räumen [5] <i>Medizin: operativ entfernen</i> [6] <i>Geometrie</i> : Strecke auf Gerade festlegen

In any event, these time-related dynamics and delimitation-related uncertainties are probably two reasons why different dictionaries contain sense inventories of different composition and detail. This is illustrated by Figure 2, which shows the sense inventory of the verb *abtragen* ‘to dismantle’ as represented by Duden<sup>9</sup> and Wiktionary.<sup>10</sup> While there are three overlaps (Wiktionary[*x*], *x* = 2, 3, 4), there is one case where a Wiktionary sense (Wiktionary[1]) is divided into two Duden senses (1.a, 1.b) and one case of senses that the other resource does not know (Wiktionary[5]) – in 2019 (download: May 1, 2019), Duden[4] was unknown to Wiktionary. While the first deviation can be seen as a difference in semantic resolution, the second raises the more fundamental question of the “true set” of different senses assumed to exist independently of scientific observation, which in turn evokes the question which of the actual senses of the verb are not “listed”. In other words, should we opt for Duden, Wiktionary, or the union of all such resources – and what does that leave open (assuming we have solved all the problems of sense matching or ontology matching as induced)?

<sup>9</sup> <https://www.duden.de/>

<sup>10</sup> <https://de.wiktionary.org/>

	Wiktionary:	Duden:
	– verbs: 14 649	– verbs: 19 278
	– senses: 29 894	– senses: 31 404
GermaNet:		
– verbs: 10 764	– same verbs: 8 440 (78%,58%)	– same verbs: 10 319 (96%,54%)
– senses: 18 336	– same sense num.: 2 798 (15%,9%)	– same sense num.: 6 120 (33%,31%)
	– same senses: 1 844 (10%,6%)	

Figure 3:  
GermaNet in relation  
to Wiktionary and Duden;  
*same verbs*:  
word-form-based counting;  
*same number of senses*:  
based on the same number  
of distinguished senses (not  
necessarily the same); *same  
senses*: based on assignable  
senses

A more systematic summary of the differences is given in Figure 3. Using version 12 of GermaNet as a reference, it shows the overlap between this resource and Wiktionary and Duden in terms of verb forms, sense numbers, and in the case of Wiktionary, senses (using the mapping between the two resources). We see both remarkably low overlaps in terms of the verbs mapped (52% of the Duden verbs are mapped by this version of GermaNet) and, even more so, in terms of the sense inventory sizes. Again, this raises the question what alignments and potential unions would be necessary to arrive at a more complete (“truer”) inventory – a task that is beyond the scope of this paper. Moreover, the first deviation in scale is related to the fact that different NLP applications require different granularities of word senses (Navigli 2009), which induces a third source of dynamics. Consequently, one might argue for an intrinsic approach that uses, e.g., transformers (Devlin *et al.* 2018) to represent senses indirectly as a result of postprocessing contextualized word representations rather than enumerating them in advance (see Pilehvar and Camacho-Collados 2021, p. 94 for an example).

While this approach has the advantage of adaptability (through fine-tuning) to ever-new corpora, it also has the disadvantage that senses appear as ephemeral entities that make identifications and comparisons across corpus boundaries difficult: ultimately, such an approach lacks a sufficient degree of explicitness necessary for delineating indisputably existing senses (see the introduction) as nameable objects of humanities research which ultimately make them a subject of separate studies. In light of these arguments, we pursue the path of using sense inventories to view word senses as *discrete, designatable*

and *nameable* entities – and see this as a kind of working hypothesis. To survey all dictionaries and sense inventories available for German is beyond the scope of this paper. Therefore we focus on frequently used resources, that is, Duden (Duden *et al.* 1980), Wiktionary (Wiktionary 2019; Mehler *et al.* 2018) and GermaNet (Hamp and Feldweg 1997; Kunze and Lemnitzer 2002; Henrich *et al.* 2012) as a taxonomy:<sup>11</sup>

1. **Duden** is a spelling dictionary of German, first published in 1880, which subdivides lemmata into senses. Duden senses are enumerated and further differentiated by enumerating more granular word senses. The feature descriptions and senses are combined with examples from German text corpora or with manually created examples. Verb entries may contain lists of synonyms, with each list roughly corresponding to one sense of the verb. However, Duden contains relations at the lemma level, not at the sense level, as the synonym lists are not connected to senses.
2. **Wiktionary** is a dictionary developed under the auspices of the Wikimedia Foundation according to the Wiki principle. Word senses are enumerated and distinguished by descriptions and examples. Wiktionary specifies relationships such as synonyms, antonyms, hypernyms and hyponyms at the sense level (but not necessarily: in some cases they are specified only at the lemma level – for the details of this model cf. Mehler *et al.* (2018)). These relations point at units at the level of superlemmas and not of senses.
3. **GermaNet** is a terminological ontology similar to WordNet (Miller 1995; Fellbaum and Miller 1998). Senses are grouped together into synsets which are networked by means of semantic relations. The GermaNet subgraph containing only verbs has a tree-like core structure based on hyponym/hypernym relations.

The choice of a sense inventory is essential to keep VSD manageable, and to be able to process corpora with existing tools or use them to extend existing corpora. GermaNet's WordNet-like structure

---

<sup>11</sup> For a lexicographic overview of web-based German dictionaries, see Storrer 2010; see Sowa 2000 for the characterization of wordnets as terminological ontologies.

Table 1: Number of verb lemmas, synsets, and senses in Duden, GermaNet and Wiktionary. Duden and Wiktionary do not (fully) specify relations at the sense level. These resources do not group senses into synsets so the corresponding entries for the number of synsets for these resources are empty. GermaNet distinguishes between senses and synsets, where the former are exemplified by sense glosses. The last row shows the coverage of the resource’s verbs by COW

	GermaNet	Duden	Wiktionary
#verb lemmas	10,764	19,278	14,649
#verb synsets	14,178	∅	∅
#senses	18,336	41,441	29,894
(senses or sense glosses)			
coverage	97.9%	93.6%	97.4%

offers many advantages for ML because of the sense relations it represents. Moreover, GermaNet describes these relations completely at the level of senses. It is constantly maintained, with several text corpora already mapped on GermaNet and tools available for their processing (Henrich and Hinrichs 2013; Henrich *et al.* 2012, 2011). Table 1 shows the number of lemmas and senses maintained by these resources: Duden contains the largest number of verbs, but the gain in coverage of the verbs annotated in COW (Schäfer and Bildhauer 2012), one of the largest openly available corpora for German, is marginal. That is, the verbs in Duden that are not included in GermaNet are apparently rare: the 9,349 verbs contained in Duden, but not in GermaNet, have a COW coverage of only 1.36%. Likewise, the 6,209 verbs contained in Wiktionary but not in GermaNet have a COW coverage of only 0.85%. Given its many advantages and its sufficiently high COW coverage, we selected GermaNet, and specifically the then current version 14, as an inventory of word senses.

### Corpus creation

### 2.2

Having decided on a verb sense inventory, the next step is to create the TTLab German Verb Sense Corpus (TGVCorp) in which a sufficiently large number of verbs from this inventory are disambiguated at the sense level. To this end, we consider three boundary conditions that

an ideal corpus should fulfill: (C1) a relevant number of verb lemmas should be covered, whose occurrences (C2) cover a large part of verb tokens observable in a reference corpus and (C3), a sufficient number of example sentences per lemma should be annotated so that ML models can be trained with this data. We choose COW as the reference corpus for C2 and use it to determine which verbs to disambiguate, and TüBa-D/Z Treebank as the text repository for examples for C3, coincidentally following the approach of Henrich (2015). This section describes how we arrive at these choices, giving an overview of existing German corpora and COW in particular in the process.

We want to prioritize high verb-token coverage (C2) over high verb-lemma coverage (C1), as this naturally helps with finding sufficient examples per lemma (C3). To do this, we process verbs according to their rank frequency distribution. This follows the idea that C2 is related to the power-law-like distribution of verb frequencies in corpora, thus selecting the most frequent verbs will quickly capture the 80% majority of verb-related tokens according to the Pareto principle (Newman 2005). In fact, the distributions of verb occurrences in a number of reference corpus candidates are heavy-tailed, see Table 2.<sup>12</sup>

Since verbs carry content as well as serve auxiliary functions, we distinguish the distribution of all verbs from that of verbs excluding modal and auxiliary verbs (that is, verbs mainly indicating possibility or necessity). The latter are usually the most frequent verbs by some distance. In order to achieve distributional profiles we compared a power law fit against a lognormal fit. Since  $R$  is negative or null in all cases, a lognormal distribution is the preferred fit. However, a lognormal fit is significant (i.e.  $p \leq 0.05$ ) only for GVSD<sup>13</sup>, Wikipedia, Gutenberg<sup>14</sup>, German Parliamentary Corpus

---

<sup>12</sup>We apply the toolbox of Alstott *et al.* (2014) according to Clauset *et al.* (2009): power laws (first) are compared to lognormal distributions (second): “ $R$  is the loglikelihood ratio between the two candidate distributions. This number will be positive if the data is more likely in the first distribution, and negative if the data is more likely in the second distribution. The significance value for that direction is  $p$ .” (Alstott *et al.* 2014, p. 5).

<sup>13</sup>German Verb Subcategorisation Database (GSDV), see Scheible *et al.* 2013.

<sup>14</sup>A free digital library with over 60,000 eBooks, including classics, for download or online reading; <https://www.gutenberg.org/>.



Table 2: Power law goodness-of-fit tests for the rank frequency distributions of verbs with and without modals (Mod.) in terms of the coefficient of (adjusted) determination (R resp.  $R^2$ ) and the Kolmogorow-Smirnow test (test value KSstat and p-value KSp)

Name	Mod.	alpha	x-min	R	P	$R^2$	Adj. $R^2$	KSstat	KSp
COW	no	2.30	1,032,974.00	-0.46	0.52	0.90	0.90	0.03	0.97
COW	yes	2.04	1,464,713.00	0.00	0.95	0.97	0.97	0.04	0.99
deCOW16B	no	2.29	819,801.00	-0.42	0.54	0.91	0.91	0.03	0.96
deCOW16B	yes	2.09	723,889.00	-0.16	0.16	0.97	0.97	0.03	0.93
DTA	no	2.12	4,567.00	-1.12	0.33	0.86	0.86	0.02	0.96
DTA	yes	2.02	4,031.00	-0.01	0.95	0.98	0.98	0.03	0.87
GVSD	no	1.50	5.00	-13.53	0.00	0.91	0.91	0.03	0.84
GVSD	yes	1.50	5.00	-12.49	0.00	0.93	0.93	0.03	0.93
Gutenberg	no	1.52	8.00	-20.03	$3.34 \times 10^{-05}$	0.91	0.91	0.03	0.67
Gutenberg	yes	1.52	8.00	-17.09	0.00	0.99	0.99	0.02	0.98
Leipzig	no	2.21	17,156.00	-1.35	0.30	0.95	0.95	0.04	0.82
Leipzig	yes	2.06	15,889.00	0.00	0.90	0.95	0.95	0.03	0.97
Parlament	no	1.40	3.00	-40.98	$4.68 \times 10^{-09}$	0.93	0.93	0.04	0.85
Parlament	yes	2.03	17,683.00	0.00	0.80	0.95	0.95	0.03	0.97
SZ	no	1.43	5.00	-50.87	$2.19 \times 10^{-11}$	0.94	0.94	0.03	0.95
SZ	yes	2.10	33,646.00	-1.04	0.14	0.96	0.96	0.02	1.00
Textbooks	no	2.24	233.00	-3.55	0.06	0.83	0.83	0.05	0.64
Textbooks	yes	2.11	219.00	-0.19	0.77	0.90	0.90	0.04	0.87
Tüba-D/Z	no	2.43	145.00	-0.33	0.58	0.92	0.92	0.03	0.99
Tüba-D/Z	yes	2.19	104.00	-1.16	0.11	0.95	0.95	0.03	0.80
Wikipedia	no	1.45	5.00	-6.81	0.01	0.81	0.81	0.04	0.54
Wikipedia	yes	1.44	6.00	-19.61	$3.28 \times 10^{-05}$	0.90	0.90	0.03	0.70
ZEIT	no	2.17	6,472.00	-0.77	0.41	0.87	0.87	0.03	0.95
ZEIT	yes	2.04	7,123.00	0.00	0.93	0.97	0.97	0.02	1.00

(GerParCor) corpus<sup>15</sup> (Abrami *et al.* 2022) and SZ<sup>16</sup> (both without modal verbs).

<sup>15</sup> A corpus of historical German parliamentary protocols from three centuries, covering four countries and processed for NLP research in political communication.

<sup>16</sup> Süddeutsche Zeitung 1992–2014

For this reason, we determined the goodness-of-fit values for fitting the distributions to a power law. Results are collected in Table 2. The (adjusted) coefficient of determination was calculated by using the curve fitting toolbox `cftool` from MATLAB (The MathWorks, Inc. 2012). The Kolmogorow-Smirnow test was carried out by using the `igraph` library (Csárdi and Nepusz 2006). The results vary from weaker fits ( $R^2 = 0.81$ ) to strong fits ( $R^2 = 0.99$ ), reflecting the distribution tests from Table 2. Furthermore, we observe no p-value smaller than 0.05 for the Kolmogorow-Smirnow goodness-of-fit test (in which case a power law distribution hypothesis would have to be rejected). Hence, although there is some distributional heterogeneity in the verb frequencies, they are nonetheless all heavy-tailed.

The question then is which of these corpora to use as a reference for determining C2. This can be answered with the help of Table 3, which shows verb token overlap among several reference corpora.<sup>17</sup> The table shows coverage of lemmas of different corpora with respect to one another, weighted by the frequency of the lemmas. A coverage of >75% is indicated by green cell color (max. ■), a coverage of <25% by red color (max. ■). Relative coverage in between (i.e., 25–75%) is colored gray (■). We treat the set of lemmas as a multiset, that is, the coverage of corpus  $A$  by corpus  $B$  for a lemma  $v \in V$  with frequency  $x_v$  in  $A$  and  $y_v$  in  $B$  is given by  $\sum_{v \in V} \min(x_v, y_v) / |A|$ , where  $|A|$  is the number of tokens in  $A$  of all lemmas in  $V$ . The number in brackets indicates the coverage of the lemmas, ignoring frequency. For a given row, the columns show how many of the lemma occurrences in that row corpus are covered by the column corpus. Note that for reference dictionaries such as GermaNet the number of occurrences per lemma is always 1 and token coverage is reduced to lemma coverage. It turns out that the largest freely available German corpus COW (Schäfer and Bildhauer 2012; Schäfer 2015), best covers all resources displayed in this heatmap. Thus we choose it as the reference for C2, selecting verbs according to their rank frequency distribution.

---

<sup>17</sup> Whenever needed, corpora were preprocessed with TextImager (Hemati et al. 2016), e.g., regarding POS tagging.

On German verb sense disambiguation

Table 3: Verb lemma frequency coverage of annotated verbs in TGVCorp with respect to German reference corpora. See Appendix B for version information

	COW	CDW16b	DeReKo (1/16)	Die ZEIT	DTA	Gutenberg	Leipziger NS	Parlament	SZ	EU Bookshop	Textbooks	Wikipedia	GVSD	Duden	wiktionary	Germanet	BabelNet	E-VALBU	TuBa-D/Z	webCade	deReC	TTVC	TTVC*
COW	—	83.4 (2.5)	17.9 (4.6)	1.1 (4.9)	1.0 (3.4)	3.3 (9.0)	2.3 (6.0)	1.5 (3.3)	4.6 (8.0)	1.7 (3.7)	0.0 (0.7)	4.2 (6.3)	1.6 (7.3)	0.0 (3.1)	0.0 (2.3)	0.0 (1.8)	0.0 (0.6)	0.0 (0.1)	0.0 (0.0)	0.0 (0.2)	0.0 (0.0)	0.0 (0.3)	0.0 (1.8)
CDW16b	100.0 (100.0)	—	21.5 (97.2)	1.3 (81.8)	1.2 (65.0)	3.9 (81.0)	2.8 (83.8)	1.8 (70.2)	5.5 (92.3)	2.0 (65.8)	0.0 (20.6)	5.0 (82.2)	2.0 (90.4)	0.0 (75.1)	0.0 (61.0)	0.0 (69.4)	0.0 (19.6)	0.0 (5.8)	0.0 (0.0)	0.0 (0.6)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
DeReKo (1/16)	98.8 (92.1)	98.6 (48.5)	—	6.0 (55.9)	5.2 (43.6)	17.3 (63.0)	12.7 (59.7)	8.4 (43.5)	25.2 (82.8)	9.0 (32.8)	0.1 (10.7)	22.0 (60.7)	9.0 (71.6)	0.0 (49.6)	0.0 (38.3)	0.0 (33.9)	0.0 (10.4)	0.0 (1.9)	0.0 (0.3)	0.0 (0.0)	0.0 (0.1)	0.0 (0.4)	0.0 (33.4)
Die ZEIT	94.7 (44.7)	94.3 (18.6)	94.4 (25.5)	—	68.5 (96.6)	87.3 (34.0)	98.3 (37.8)	80.0 (28.5)	94.7 (37.1)	68.3 (34.0)	1.4 (5.5)	85.8 (30.6)	90.2 (32.9)	0.1 (29.6)	0.1 (16.8)	0.1 (15.0)	0.0 (4.7)	0.0 (6.9)	0.1 (0.1)	0.0 (1.5)	0.0 (0.0)	0.3 (0.0)	0.1 (64.8)
DTA	92.2 (46.8)	91.5 (17.5)	90.5 (23.6)	75.1 (31.5)	—	91.1 (46.1)	86.5 (30.3)	71.4 (23.5)	87.4 (32.5)	62.9 (14.1)	1.5 (0.5)	79.2 (31.4)	78.3 (31.6)	0.1 (22.6)	0.1 (17.1)	0.1 (15.4)	0.0 (5.1)	0.0 (1.0)	0.1 (0.2)	0.0 (1.8)	0.0 (0.0)	0.3 (2.8)	0.1 (15.2)
Gutenberg	95.0 (40.3)	94.1 (81.1)	91.7 (25.5)	29.2 (16.8)	27.7 (19.2)	—	50.5 (16.6)	34.0 (12.0)	77.1 (6.7)	28.8 (2.9)	0.5 (2.9)	56.1 (26.7)	38.4 (25.1)	0.0 (11.8)	0.0 (6.8)	0.0 (7.3)	0.0 (2.8)	0.0 (6.4)	0.0 (0.1)	0.0 (0.8)	0.0 (0.1)	0.1 (0.2)	0.0 (67.2)
Leipziger NS	94.9 (66.6)	94.7 (23.5)	94.7 (33.5)	46.4 (46.6)	37.2 (31.6)	71.3 (41.4)	—	56.1 (34.2)	94.5 (17.3)	47.6 (7.7)	0.7 (39.8)	77.2 (39.8)	60.0 (44.0)	0.1 (23.8)	0.0 (20.7)	0.0 (5.9)	0.0 (1.1)	0.0 (0.2)	0.0 (1.0)	0.0 (0.0)	0.1 (3.0)	0.1 (18.4)	
Parlament	94.2 (62.7)	94.0 (34.0)	93.8 (42.3)	56.5 (60.7)	45.9 (42.3)	71.8 (51.7)	84.0 (59.2)	—	89.5 (30.8)	68.0 (11.3)	1.0 (49.0)	74.5 (49.0)	70.4 (62.8)	0.1 (35.6)	0.1 (28.7)	0.1 (9.4)	0.0 (4.9)	0.1 (0.3)	0.1 (3.2)	0.0 (0.1)	0.2 (5.2)	0.1 (27.8)	
SZ	98.0 (31.7)	97.7 (14.9)	97.8 (24.8)	23.2 (26.3)	19.5 (19.5)	56.6 (41.1)	49.1 (29.3)	31.1 (19.3)	—	30.2 (10.4)	0.4 (4.7)	62.0 (40.1)	34.7 (38.9)	0.0 (17.3)	0.0 (13.3)	0.0 (11.3)	0.0 (3.5)	0.0 (0.6)	0.0 (0.1)	0.0 (1.1)	0.0 (1.8)	0.0 (1.8)	
EU Bookshop	100.0 (100.0)	100.0 (100.0)	98.8 (99.9)	98.8 (93.4)	39.9 (79.4)	59.9 (90.5)	70.2 (93.9)	67.0 (88.0)	85.8 (97.4)	—	1.0 (31.1)	83.7 (95.0)	66.7 (97.1)	0.0 (88.9)	0.0 (75.6)	0.0 (79.3)	0.0 (27.9)	0.0 (5.8)	0.1 (0.9)	0.0 (10.0)	0.0 (0.2)	0.2 (16.3)	0.0 (78.3)
Textbooks	94.5 (47.7)	88.1 (47.7)	88.2 (49.5)	89.5 (56.2)	85.3 (48.7)	88.4 (59.8)	89.8 (58.2)	89.6 (58.0)	90.5 (47.3)	88.0 (47.3)	—	90.7 (64.8)	89.7 (65.1)	1.5 (48.3)	1.4 (45.4)	1.5 (47.1)	0.0 (2.4)	0.0 (0.7)	2.7 (1.3)	1.1 (11.9)	0.0 (0.2)	1.3 (19.0)	1.4 (46.6)
Wikipedia	97.7 (42.1)	97.7 (13.9)	93.0 (20.5)	22.9 (22.7)	19.3 (19.7)	44.8 (40.1)	43.7 (24.0)	28.2 (17.1)	67.5 (41.9)	32.1 (10.5)	0.4 (4.7)	—	34.0 (33.3)	0.0 (16.9)	0.0 (13.1)	0.0 (3.7)	0.0 (0.6)	0.0 (0.1)	0.0 (1.1)	0.0 (0.0)	0.1 (1.8)	0.0 (1.4)	
Duden	97.4 (42.9)	96.7 (13.5)	96.9 (21.1)	61.7 (21.3)	48.8 (17.3)	78.6 (33.0)	87.0 (23.1)	68.1 (16.0)	96.9 (35.4)	65.6 (9.4)	1.0 (4.1)	87.2 (29.0)	—	0.1 (15.2)	0.1 (11.7)	0.1 (10.2)	0.0 (3.2)	0.0 (0.6)	0.1 (0.1)	0.0 (0.0)	0.0 (1.6)	0.1 (10.0)	
Germanet	90.8 (89.8)	55.8 (55.8)	73.8 (73.8)	67.4 (67.4)	62.3 (62.3)	78.3 (78.3)	68.4 (68.4)	54.5 (54.5)	79.4 (79.4)	43.4 (43.4)	15.5 (15.5)	74.4 (74.4)	76.4 (76.4)	—	61.9 (61.9)	51.5 (51.5)	15.8 (15.8)	2.9 (2.9)	0.4 (0.4)	4.9 (4.9)	0.1 (0.1)	8.0 (8.0)	50.8 (50.8)
BabelNet	89.8 (89.8)	59.6 (59.6)	75.1 (75.1)	70.1 (70.1)	62.1 (62.1)	76.6 (76.6)	72.4 (72.4)	57.8 (57.8)	80.7 (80.7)	48.6 (48.6)	19.2 (19.2)	75.9 (75.9)	77.8 (77.8)	81.5 (81.5)	—	57.6 (57.6)	20.0 (20.0)	3.8 (3.8)	0.6 (0.6)	6.5 (6.5)	0.1 (0.1)	10.3 (10.3)	56.7 (56.7)
E-VALBU	95.3 (69.3)	80.3 (80.3)	90.3 (90.3)	87.9 (87.9)	76.2 (76.2)	86.9 (86.9)	88.6 (88.6)	77.2 (77.2)	93.4 (93.4)	69.3 (69.3)	27.0 (27.0)	90.9 (90.9)	91.9 (91.9)	92.2 (92.2)	78.4 (78.4)	—	26.5 (26.5)	5.2 (5.2)	0.8 (0.8)	8.9 (8.9)	0.1 (0.1)	14.5 (14.5)	98.6 (98.6)
TuBa-D/Z	69.3 (69.3)	61.4 (61.4)	65.3 (65.3)	65.2 (65.2)	59.9 (59.9)	64.1 (64.1)	65.7 (65.7)	60.6 (60.6)	68.0 (68.0)	57.5 (57.5)	31.7 (31.7)	68.0 (68.0)	67.2 (67.2)	66.6 (66.6)	64.1 (64.1)	62.6 (62.6)	—	10.2 (10.2)	1.4 (1.4)	15.0 (15.0)	0.3 (0.3)	20.3 (20.3)	61.9 (61.9)
webCade	96.1 (96.1)	95.8 (95.8)	95.9 (95.9)	96.1 (96.1)	95.8 (95.8)	96.1 (96.1)	96.1 (96.1)	96.1 (96.1)	96.8 (96.8)	84.0 (84.0)	96.1 (96.1)	96.1 (96.1)	96.1 (96.1)	96.8 (96.8)	98.4 (98.4)	98.6 (98.6)	82.2 (82.2)	—	7.6 (7.6)	55.0 (55.0)	2.1 (2.1)	67.0 (67.0)	98.1 (98.1)
deReC	100.0 (99.8)	100.0 (99.6)	100.0 (99.9)	100.0 (99.8)	99.6 (99.8)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.7)	100.0 (99.8)	100.0 (99.7)	100.0 (99.8)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.9)	100.0 (99.8)
TTVC	100.0 (100.0)	99.9 (99.6)	100.0 (99.9)	99.8 (99.6)	99.1 (99.7)	99.7 (99.9)	99.8 (99.9)	99.7 (99.9)	99.8 (99.8)	99.6 (99.8)	67.7 (75.2)	100.0 (99.8)	100.0 (100.0)	1.0 (99.2)	3.9 (96.6)	4.0 (99.7)	2.4 (59.5)	1.0 (24.4)	0.5 (0.9)	5.1 (41.0)	0.0 (1.0)	—	4.0 (99.7)
TTVC*	95.3 (95.3)	80.3 (80.3)	90.4 (87.9)	87.9 (87.9)	76.0 (76.0)	86.9 (86.9)	88.6 (88.6)	77.3 (77.3)	93.4 (93.4)	69.4 (69.4)	27.1 (27.1)	90.8 (90.8)	91.9 (91.9)	92.2 (92.2)	78.3 (78.3)	100.0 (100.0)	26.6 (26.6)	5.2 (5.2)	0.7 (0.7)	9.0 (9.0)	0.1 (0.1)	14.7 (14.7)	—

COW is a web-crawled corpus containing 807,782,354 sentences. Due to its automatic pre-processing, it contains a considerable number of lemmatization and POS tagging errors. This explains the unusually high number of verb lemmas found in COW (see Table 4). To fix these errors, we apply four heuristics to the selection of verb lemmas output

Table 4:  
COW-based statistics  
of verb lemmas  
and their tokens

	Plain	Filtered
# verb lemmas	368,677	41,316
# verb tokens	939,732,595	880,670,918
% verb hapax legomena	50%	35%

by the lemmatization of COW:

1. The lemma candidate must be in present infinitive and thus end in *-n*.
2. It has to consist of at least 2 characters.
3. It must be in lower case.
4. Modal and auxiliary verbs are excluded.

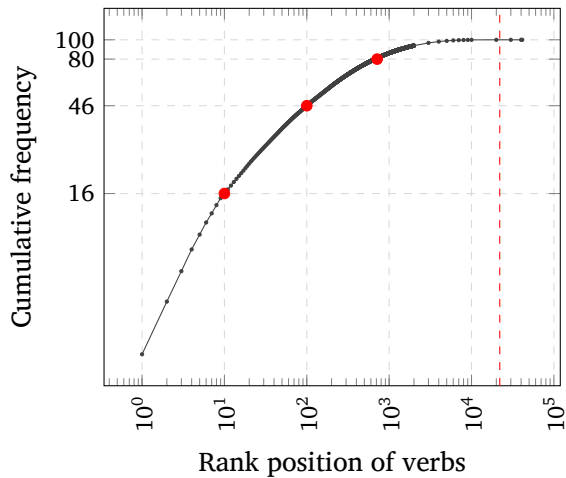
Using these heuristics, 88 % of verb lemmas in COW are removed, but only 6 % of verb tokens (see Table 4).

The frequencies of the remaining verb lemmas are plotted in Figure 4 as a cumulative rank frequency distribution.

We observe that a small number of verbs covers a large number of verb tokens. More specifically, the 945 most frequent verbs cover 80 % of COW’s verb tokens. A corpus disambiguating a sufficient number of examples for each of these lemmas would thus satisfy C2 and C3.

However, not all of these verbs are ambiguous, and some have already been annotated. And while we prioritize C2 over C1, we would

Figure 4:  
The cumulative  
distribution of the token  
frequencies of the verbs  
in the COW corpus. The 945  
most common verb lemmas  
cover 80 % of the verb  
tokens in COW



still like to satisfy C1 to the largest degree allowed by our resources. Thus, we select verbs to disambiguate, in descending order of their frequency according to the following criteria:

1. The lemma candidate has at least two senses in GermaNet.
2. It is not already annotated in TüBa-D/Z.
3. It is not a modal verb and not an auxiliary verb.

The result is a set of 1,560 ambiguous verbs with a COW coverage of 78%.

The third condition, C3, concerns the selection of a corpus to be sense-annotated based on our reference set of verbs. Here we started from TüBa-D/Z, a German newspaper corpus, which is annotated semi-automatically at several linguistic levels (Telljohann *et al.* 2012). Parts of TüBa-D/Z are also already sense-annotated. We thus “filled out” an existing corpora instead of starting from scratch.

We also added sentences from other resources to fill in gaps in lemma coverage. More specifically, we included sentences from E-VALBU and the SALSA 2.0 Corpus (Burchardt *et al.* 2006) that are linked to semantic annotations in Berkeley FrameNet (Ruppenhofer *et al.* 2016) format. In this way, future work will gain access to relations between verb-related frames and the verb senses we annotate.

TGVCorp is thus generated as a union of three corpora: TüBa-D/Z, SaLsa and E-VALBU – see Table 5 for the corpus statistics. Multiple

Sources	TüBa-D/Z, SaLsa, E-VALBU
Total # of sentences	31,650
Total # of annotated word lemmas	1,560
Total # of tagged word tokens	39,241
Frequency range (occurrences/lemma)	1–261
Average frequency (occurrences/lemma)	25
Polysemy range in GermaNet (senses in GermaNet/lemma)	1–26
Average polysemy in GermaNet (senses in GermaNet/lemma)	3.27
Polysemy range of occurring words (occurring senses/lemma)	1–18
Average occurring polysemy of lemmas (occurring senses/lemma)	2.34
Average occurring polysemy of words (occurring senses/word)	3.77

Table 5:  
TGVCorp  
breakdown

Table 6:  
Verb lemmas and  
tokens in various  
corpora and their  
coverage with  
respect to COW

	TüBa-D/Z	WebCAGE	deWaC	TGVCorp
# verb lemmas	82	959	15	1,560
# verb tokens	9,290	3,186	608	39,241
average frequency	113	3	41	25
average polysemy	2.5	3.7	7.9	2.34
COW coverage (lemma-based)	6.2%	66.4%	6.4%	78.02%

other corpora are also annotated with GermaNet senses. These are the sense-annotated sections of TüBa-D/Z itself, WebCAGE (Henrich *et al.* 2012) and deWaC (Raileanu *et al.* 2002). Table 6 compares our target corpus to these, demonstrating that only TGVCorp offers a high COW coverage with a large number of lemmas and at the same time a sufficiently high number of example sentences per lemma. This closes the gap left by its competitors.

### 2.3

#### *Annotating TGVCorp*

We developed `VerbSenseAnnotator`<sup>18</sup> to disambiguate TGVCorp at the sense level, and conducted this annotation in two stages. As in related approaches (Henrich 2015; Kilgarriff 1998; Fellbaum *et al.* 2001; Saito *et al.* 2002; Passonneau *et al.* 2012), `VerbSenseAnnotator` shows sentences in which the occurrences of target verbs are to be disambiguated on the level of lemmas. Sentences are preprocessed by `TextImager` to capture lemma, POS, and dependency structure information, and to present verbs with corresponding senses from GermaNet. For each target sense of each target verb, the corresponding synonyms, hyponyms, and hypernyms are listed, as well as sense descriptions and example sentences where available, so that annotators can disambiguate more easily. Ideally, exactly one meaning should be selected for each occurrence of each target verb, but when in doubt, more than one is possible. Occurrences of target verbs for which the annotator cannot find a sense in `VerbSenseAnnotator` can be marked. If multiple senses or no appropriate sense are selected for

<sup>18</sup><https://textimager.hucompute.org/VSD/>

a verb occurrence, this indicates that the verb's sense definitions are problematic. Commonly, this was a problem with very fine-grained sense definitions, which are indistinguishable for annotators that have to rely on short sense descriptions and example sentences. Other problematic cases were metaphorical usages or hierarchical senses, such as *laufen* in the sense of movement on foot in general, 'to move' vs. *laufen* in the sense of a fast, running movement, 'to run'. Following the approach of Palmer *et al.* (2007), these senses with very low inter-annotator agreement were manually reviewed and merged if required. A list of all senses merged in this fashion is shown in Appendix A.

To evaluate the quality of verb-sense annotation, each target sentence was annotated independently by several annotators in two stages. The first stage comprised the bulk of annotation work, in which a total of 19 annotators participated, including undergraduates, graduate students, doctoral students, and postdoctoral fellows in computer science and computational linguistics. The second stage involved 7 annotators. The procedure was the same for both stages, with two exceptions. The first difference was in the choices annotators had. In the first stage, they could select multiple senses for a single instance. This was not possible in the second stage, where the annotators had to select a single sense. In addition, they could mark sentences that were ambiguous or incomprehensible due to a lack of context. The second difference relates to the selection of the gold label in situations where annotators disagreed. To address this issue during the first stage, we developed a method that compares the inter-annotator agreement between each annotator and the original TüBa-D/Z annotation to prefer the annotator with the highest agreement.<sup>19</sup> Therefore, in order to be consistent with the TüBa-D/Z interpretations, we decided to prefer the annotator who agreed in the majority of cases. Given this approach, we do not know with certainty the reliability of our annotations. However, by selecting the annotator this way, and manually checking senses with low agreement between annotators, we guarantee at least a strong orientation towards TüBa-D/Z, even if this is certainly not the only authoritative resource. In the second

---

<sup>19</sup>This approach is motivated by the fact that annotators often agreed on the distinction of senses, but not on their interpretations (i.e. they agreed that a verb has *n* different senses, but not on what these senses are).

stage, each disagreement was checked and a gold label was manually selected. During this process, we discovered many senses with very low inter-annotator agreement.

### 3 A SIMPLE METHOD FOR AUTOMATIC VSD

Using TGVCorp, we train a supervised system for VSD by elaborating the approach of Hemati (2020). We follow approaches that use human-annotated training data to learn to assign senses from predefined lexical resources to ambiguous lexical text occurrences (Hemati 2020; Henrich 2015; Papandrea *et al.* 2017; Luo *et al.* 2018; Peters *et al.* 2018; Melamud *et al.* 2016; Uslu *et al.* 2018). One of the most elaborate early approaches to WSD in German is that of Henrich (2015), who uses GermaNet as a sense inventory to train supervised and knowledge-based systems. A problem faced by these and related approaches is that the underlying annotated corpora usually only contain a few lemmas or have very few annotated instances per lemma. Although TGVCorp is one step ahead in filling this gap, sense compression must be performed for tackling the latter bottleneck, as will be explained below. To perform VSD, we train TTvSense, a supervised classifier based on fastSense (Uslu *et al.* 2018), which in turn is based on fastText (Joulin *et al.* 2017; Bojanowski *et al.* 2016). TTvSense is a feed-forward network that includes sense compression according to Vial *et al.* 2019. We compare TTvSense with EWISER (Bevilacqua and Navigli 2020), a state-of-the-art approach to WSD, and show how to circumvent the data bottleneck problem in VSD using language models. To compare EWISER and TTvSense, we reproduce the method of Henrich (2015) using the *TüBa-D/Z Gold Standard for Supervised WSD* corpus, focusing on verbs (see Table 7 for its statistics). We split this data to maintain the following ratio per lemma (Henrich 2015; Botev and Ridder 2017; Witten *et al.* 2011): 60% for training, 20% for validation and 20% for testing. For methods that do not require validation sets, this part was omitted to keep training and test sets comparable.



	GermaNet	WordNet Subset
Total # of annotated word lemmas	82	68
Total # of tagged word tokens	9,290	5,765
Frequency range (occurrences/lemma)	1–822	2–280
Average frequency (occurrences/lemma)	113.3	84.8
Polysemy range in GermaNet (senses in GermaNet/lemma)	1–14	—
Average polysemy in GermaNet (senses in GermaNet/lemma)	2.9	—
Polysemy range of occurring words (occurring senses/lemma)	1–9	1–4
Average occurring polysemy of lemmas (occurring senses/lemma)	2.45	1.74
Average occurring polysemy of words (occurring senses/word)	3.16	1.97

Table 7:  
TüBa-D/Z  
sense annotation  
subset for  
supervised WSD  
Henrich (2015),  
verbs only

### TTvSense

3.1

TTvSense represents a word as a sum of  $n$ -gram vectors, where the word itself is one of the  $n$ -grams initialized from previously trained word embeddings. These word representations are fine-tuned during the training. A sentence is encoded by averaging the word representations for all words contained in it. This sentence encoding forms the input for a single fully connected layer, which produces output scores for all senses of all lemmas. Finally the output senses are filtered to remove all which do not belong to the current target lemma. The list of valid senses for the target lemma is obtained from the training corpus as part of the training process. To extend this model, we performed sense compression on GermaNet according to Vial *et al.* (2019). In this process, all senses for a given lemma are removed from their original synset and reassigned to be just below the last common ancestor in the hyperonymy hierarchy. The procedure is explained in detail in Section 3.5.

TTvSense uses information about the target word only after the scores have been calculated. Furthermore, it does not process posi-

tion or word order information. This is a problem when a sentence manifests several disambiguation-relevant contexts due to its clause structure. For example, the first half of the sentence *Er lief ins Büro und machte den Rechner an*. ‘He ran into the office and turned on the computer’ indicates a motion sense of *lief* ‘ran’ that is not matched by the second half which might indicate another sense of that verb (*Der Computer lief* ‘The computer was running’). Without position and target information, the classifier cannot distinguish these contexts, thus accuracy suffers. To deal with this problem, we split sentences along conjunctions and punctuation marks and processed only the segment that contained the target word.

## 3.2

## EWISER

EWISER (Bevilacqua and Navigli 2020) sums the last four layers of BERT (Devlin et al. 2018) and normalizes them to a context vector  $H_0$ , which is fed into a two-layer fully-connected network to produce output values  $Z$ :

$$H_1 = \text{swish}(H_0W + b)$$

$$Z = H_1O$$

The first layer is a traditional, fully connected layer with a Swish (Ramachandran et al. 2017) activation function and is used to re-encode  $H_0$  from BERT to have the same dimensionality as the pretrained sense embeddings  $O$ . The weights of the second layer are initialized with  $O$  to produce logits for each sense in the inventory. Finally, these logits are modified based on the graph structure of the given WordNet to produce “structured logits”. For a given synset  $s$  with logit  $z_s$  and  $n_s$  related synsets  $z_i$  a new structured logit  $q_s$  is computed by adding the logits of all related synsets:  $q_s = z_s + \sum_i z_i/n_s$ . This takes the form of a residual layer where the weights are initialized by an adjacency matrix  $A$  in which the entries of each row sum up to 1:

$$Q = ZA^T + Z$$

During training the underlying BERT model is kept frozen while the weights  $A$  are fine-tuned. The sense embeddings follow a freeze-and-thaw training scheme where they are kept frozen for the first  $n$  epochs before being unfrozen and fine-tuned during the remaining epochs.

We conducted a series of experiments with German and English data and performed comparisons on English verbs from Navigli *et al.* (2017). Since EWISER requires WordNet or BabelNet (Navigli and Ponzetto 2012) labels, we experimented on the subset of TüBa-D/Z for which there are mappings from GermaNet to WordNet. The experiments are repeated for TGVCorp. The GermaNet senses in texts were mapped to WordNet using EuroWordNet’s (Vossen 1998) Inter-Lingual Index. This mapping is not complete and does not ensure a one-to-one relation, so we removed all instances for which there is no mapping. In cases with multiple relevant labels we only considered the first one provided by the mapping, discarding any others. The resulting WordNet subset is considerably smaller than the original corpus, with fewer examples per lemma and significantly lower polysemy. See Table 7 above for a comparison. The mapping from WordNet to BabelNet is done in EWISER itself, but requires updating multiple dictionary files. EWISER operates only on a subset of the BabelNet-WordNet mapping that matches entries in these files. These dictionaries limit the lemmas and the labels for each lemma which the system will produce. The pretrained checkpoint comes with multilingual dictionaries based on SemEval tasks. Testing the pretrained checkpoint on TüBa-D/Z, EWISER achieves only 53% with these dictionaries, 69% if we update the dictionaries to include the labels in the test set, and 78% if we additionally remove all labels which do not occur in the test set. Accurate dictionaries are critical to achieving good results in practice.

For EWISER we tested three different models. One was trained only on the training section of TüBa-D/Z and one on both the TüBa-D/Z training section and the WordNet Glosses and Examples corpora. Due to time and computational restraints we chose the best performing hyperparameters from Bevilacqua and Navigli 2020 for training. We also tested the pretrained multilingual model provided by Bevilacqua and Navigli 2020.

For TTvSense we examine the impact of the sentence fragmentation and sense compression over the baseline classifier. Hyperparameters were optimized on the validation set of TüBa-D/Z using Tree-structured Parzen Estimator (TPE) (Bergstra and Bengio 2012)

Table 8:  
Hyperparameters  
of training  
TTvSense

Epochs	40
Initial learning rate	0.2
Hidden dim	100
Window size	3
Loss	softmax
Pretrained embeddings	Mikolov embeddings computed by means of the Süddeutsche Zeitung corpus (1992–2014)

Table 9: EWISER hyperparameters. Training takes place in two stages where the sense embeddings are kept frozen during the first stage and fine-tuned during the second

Epochs first stage	50
Epochs second stage	20
Initial learning rate first stage	$10^{-4}$
Initial learning rate second stage	$10^{-5}$
BERT model	bert base multilingual cased
Hidden dim	512
Sense embeddings	SensEmbBERT + LMMS
Structured logits	hypernyms, derivational, verb group, similarity

as implemented by hyperopt (Bergstra *et al.* 2013). The hyperparameters for TTvSense and EWISER are shown in Tables 8 and 9.

Both EWISER and our classifier use dictionaries to limit output senses for each lemma. These essentially form another hyperparameter. For our experiments, these dictionaries were computed before the training process, excluding all senses that did not appear in the training corpora. Results are shown in Table 10. We outperform EWISER in all German tests, but perform significantly worse on the English corpora. However, our fastText-based classifier trains and evaluates much faster despite not using a GPU. Training on our machine with an AMD FX-8350 and GTX 1070 on TüBa-D/Z only, our classifier took about 4 minutes on the CPU, while EWISER took about 30 minutes despite also using the GPU. This is repeated during evaluation, with TTvSense evaluating the entire test set in less than one second, compared to about 45 seconds for EWISER. In times of problematic CO<sub>2</sub> emissions by NLP (Bender *et al.* 2021), this is a relevant finding.

Table 10: VSD results on TüBa-D/Z sense annotation subset for supervised WSD. For EWISER the subscripts indicate the source/training corpora. For TTVSense the subscripts indicate sentence fragmentation (sf) and sense compression (sc)

System	Base Corpus	Micro F1 score
Most frequent sense		71.75
Context2Vec		76.04
Best of Henrich (2015)	TüBa-D/Z with	80.74
Flair	GermaNet Labels	83.13
TTvSense		80.93 ± 0.39
TTvSense <sub>sf</sub>		87.39 ± 0.81
TTvSense <sub>sf+sc</sub>		89.14
Most frequent sense		87.24
EWISER <sub>tueba</sub>		88.43 ± 0.63
EWISER <sub>tueba + WNGC</sub>		90.94 ± 0.37
EWISER <sub>multilingual pretrained</sub>	WordNet subset	78.13
TTvSense	of TüBa-D/Z	88.79 ± 0.14
TTvSense <sub>sf</sub>		93.13 ± 0.85
TTvSense <sub>sf+sc</sub>		93.52 ± 0.29

Table 11: VSD results on SemCor and SENSEVAL

System	Micro F1 score
TTvSense <sub>sc</sub>	43.91
TTvSense <sub>sf+sc</sub>	46.94
TTvSense <sub>sf+sc</sub> on SemCor only	55.67
EWISER	69.40

We also ran comparisons on English verbs using SemCor (Miller *et al.* 1994; Navigli *et al.* 2017) as training data and the concatenation of English WSD SENSEVAL tasks as test data. We tried to determine generalization errors of our classifier by also training and testing on SemCor verbs only, using the same splitting as for TüBa-D/Z. The results are shown in Table 11 and discussed below. We then tested TTVSense on TGVCorp. The results are shown in Table 12.

Table 12:  
VSD results on TGVCorp

System	Micro F1 score
TTvSense	63.2 ± 0.4
TTvSense <sub>sf</sub>	69.8 ± 0.1
TTvSense <sub>sf+sc</sub>	65.5 ± 0.2

### 3.4

#### *Discussion*

TTvSense outperforms EWISER on both TüBa-D/Z and TGVCorp, even when taking the WordNet Gloss Corpus as additional training data for EWISER. Interestingly, this result is not repeated in English, where our classifier performs much worse. We think that this could be due to two main factors: In the German experiments, we obtained training and test data from TüBa-D/Z based on a single newspaper. SemCor, on the other hand, is based on the Brown Corpus, which contains various newspapers, books, and other sources. SENSEVAL comes mainly from articles in the Washington Post. The improvement when testing and training only on SemCor might indicate that our classifier overfits on the training data and generalizes worse than EWISER. At the same time, the increase is too small to explain the whole performance gap between German and English. The second effect is language-specific. Our classifier uses averaged word form embeddings as the context vector. This approach might work better for German than for English, since the morphology in German is more extensive, reducing the importance of positional information. However, positional information is still relevant due to sentence-internal contexts belonging to different verbs. TTvSense reflects this through its simple sentence segmentation algorithm, which performs worse on English data due to different punctuation rules. The sentence segmentation reduces error rates by around a third in all German tests, but only by about 5% in English tests. In any case, TTvSense, which we trained to disambiguate 1,560 German high-relevance verbs (see above), is a classifier for VSD that represents a new state of the art for German verbs.

### 3.5

#### *An experiment in sense compression*

Supervised systems rely on annotated training data and cannot directly disambiguate senses which they have not seen. Sense compression is

a method of extending the coverage of existing annotations by exploiting the hyperonymy structure. For this, we adapt the algorithm of Vial *et al.* (2019) for GermaNet. We consider GermaNet as a graph  $G = (V, E)$ , where the set of vertices consists of synsets  $S$  and senses (GermaNet LexUnits)  $L$  with  $V = S \cup L$  and

$$(1) \quad E = \left\{ (u, v) : \begin{array}{l} (u, v \in S, u \text{ is hypernym of } v) \\ \vee (v \in S, u \in L, u \text{ is member sense of } v) \end{array} \right\}$$

$G$  is directed and acyclic, where each vertex in  $L$  is a leaf node and only vertices in  $L$  are leaves. Using  $G$ , a graph variant  $G'$  is created as follows: pick a lemma  $v$  and select the set of vertices

$$(2) \quad L_v = \{l \in L : l \text{ belongs to lemma } v\}$$

which corresponds to the set of senses which belong to lemma  $v$ . Then mark all vertices which are ancestors of more than one  $l \in L_v$ . Finally, add an edge for every  $l \in L_v$  between  $l$  and the child of its first marked ancestor and remove the edge between  $l$  and its original synset. This ensures that only one sense per lemma per synset exists without violating the hyperonymy structure of the graph. Repeat this process for every lemma. Finally, remove any synsets that do not have any attached senses.

For a given sense  $l \in L$  the new label is determined by its direct parent. Given a target lemma and a compressed synset  $s$  one can convert back to the original sense label by searching the direct children of  $s$  for the one sense belonging to the target lemma. This procedure – see Algorithm 1 – guarantees that each synset contains only one sense per lemma, provided that the original graph fulfills the same condition. The statistics for Algorithm 1 operating on GermaNet are listed in Table 13. To quantify the effectiveness of sense compression, we performed an out-of-sample test by removing lemmas from the dataset such that there were at least 10 training instances left for each of the compressed synsets. The instances belonging to the removed lemmas formed the test set. Note that synsets can have less than 10 training instances, in which case the associated lemmas are not taken into account for removal. The results for this test are shown in Table 14.

This out-of-sample test shows that we achieve about 60% F1 score on TGVCorp (ca. 70% on TüBa-D/Z) from scratch with the compression algorithm – the alternative, of course, would be 0%.

```

Algorithm 1: for each verb  $v$  do
  Algorithm for sense compression
  | /* Mark descendants of more than one sense */
  | for each vertex  $l$  in  $L_v$  do
  |   while  $l$  is not null do
  |     if  $l.mark$  is not 'unmarked' then
  |       |  $l.mark = 'conflict';$ 
  |     else
  |       |  $l.mark = 'visited';$ 
  |     end
  |      $l = \text{parent of } l;$ 
  |   end
  | end
  | /* Reattach senses */
  | for each vertex  $l$  in  $L_v$  do
  |    $current = l;$ 
  |   while mark of parent of  $current$  is not 'conflict' do
  |     |  $current = \text{parent of } current;$ 
  |   end
  |   Remove edge between  $l$  and parent of  $l;$ 
  |   Add edge between  $l$  and  $current;$ 
  | end
  | end
  | /* Cleanup of empty synsets */
  | for vertex  $v$  in  $S$  do
  |   if  $v$  has no children in  $L$  then
  |     | Reattach children of  $v$  to parent of  $v;$ 
  |     | Remove  $v$  from graph;
  |   end
  | end
end

```

Table 13:  
Results  
of compressing  
GermaNet

	Pre-compression	Post-compression
# Synsets	14,179	1,633
Average # senses per synset	1.29	11.89
Average depth of senses	6.71	2.85
Highest depth	16	14



	TGVCorp	TüBa-D/Z
F1 Score	60.62 ± 0.69	69.53 ± 0.18
Size of train set	≈ 18700	≈ 6000
Size of test set	≈ 17500	≈ 3100
# Lemmas removed	803	37/38

Table 14:  
Results  
for the out-of-sample tests  
using the sense  
compression algorithm

### Trying to leverage language models

3.6

WSD is challenged by the data bottleneck problem (Navigli 2009). We attempt to address this problem beyond costly annotation by using language models (Devlin *et al.* 2018) that can be fine-tuned for downstream tasks (Zhou and Srikumar 2022) – here language generation (Rothe *et al.* 2020). That is, we use BERT (Devlin *et al.* 2018) to extend TGVCorp by generating new sentences starting from manually annotated ones. Following Ravfogel *et al.* (2020), we iteratively mask and replace words in sentences from left to right by sampling from the top  $k$  suggestions provided by BERT. Unlike Ravfogel *et al.* (2020), we do not only sample content words like nouns. German is less analytical than English, so substituting nouns alone easily leads to ungrammatical sentences due to agreement errors. We address this issue by processing sentences in two passes. In the first pass, nouns, adjectives, substitution pronouns, and adverbial adjectives are substituted; in the second pass, all other words are processed, leaving annotated verbs and punctuation untouched. Note that we do not try to maintain the POS of the source word, nor the original number of BERT tokens. For words consisting of multiple WordPiece tokens (Wu *et al.* 2016), we mask all tokens and replace them from left to right. To minimize morphological inconsistencies, however, only the first of them is sampled using BERT and then the top suggestions are selected for the remaining tokens (dependent selection). For example, after replacing the first token in “Schaff ###ner” with “Kell [MASK]”, the only viable option for “###ner” is identity substitutions; if this were excluded and one were to sample independently from the top  $k$  BERT suggestions, the result would likely be a non-word. The whole procedure serves to ensure both semantic variability and a certain degree of grammatical correctness. Table 15 exemplifies our procedure.

Table 15: Left: Source sentences in which words to be replaced are in italics. Right: sentence candidate in which the italicized word is predicted by BERT for the masked word in the source sentence

Source sentence	Generated sentence candidate
Der <i>Schaffner</i> läuft zum <i>Bahnhof</i> .	Der <i>junge Mann</i> läuft zum <i>Flughafen</i> . Der <i>Bursche</i> läuft zum <i>Metzger</i> . Der <i>Fünfjährige</i> läuft durchs <i>Tor</i> .
Die <i>Diskussion</i> hat mein <i>Denken</i> zu diesem <i>Thema</i> verändert.	Die <i>Diagnose</i> hat mein <i>Vertrauen</i> zu dem <i>Institut</i> verändert. Die <i>Vergangenheit</i> hat meine <i>Einstellung</i> zu dem <i>Job</i> verändert. Die <i>Debatte</i> hat mein <i>Fazit</i> zu meinem <i>Amt</i> verändert.
Das <i>Gerät</i> läuft <i>einwandfrei</i> .	Das <i>Program</i> läuft <i>jetzt bis 2020</i> . Das <i>Geschäft</i> läuft <i>im Moment gut</i> . Das <i>Haus</i> läuft <i>immer noch leer</i> .

Table 16:  
F1 scores when training our classifier  
with additional sentences from BERT.  
Baseline score is 87.3%

	$k$	3	30	100
	1	86.3	86.4	86.0
$n$	3	85.9	85.7	85.4
	10	—	84.1	83.9

We evaluate this approach of generating new, similar sentences from annotated seed sentences, by extending TüBa-D/Z using this method and training TTvSense on the new training data. We have two new hyperparameters in this approach: (1) the number of new sentences  $n$  for each seed sentence and (2) the depth  $k$  to which we sample content words. Only sentences from the training subset were selected as seed sentences. We trained with sentence fragmentation but without sense compression. The results are shown in Table 16.

It is obvious that forming new sentences in this way did not improve the results. The reason could be that our sentence generator interpolated only in the range of sentence patterns already observed in the training corpus, introducing errors that made training more difficult. While this is disappointing in light of increasingly better and

more diverse text generators, it points to a general problem of poor extrapolation capabilities of such approaches, which requires far more research to overcome. Although scores did not improve they also did not meaningfully degrade even with deep sampling. This suggests that this method could be used to create “look-alike” corpora.

### Optimising TTvSense for VSD on TGVCorp

3.7

This section explains how TTvSense was optimized for TGVCorp. Since it is a sequence classifier that does not receive information about the target lemma, TTvSense has difficulties with longer sentences. To improve it, the aforementioned sentence segmenter was used in both training and testing. Table 17 shows that it improves VSD significantly.

	TüBa-D/Z	TGVCorp
w/o splitting	78.97 %	62.07 %
with splitting	86.16 %	71.38 %

Table 17:  
Micro-F1 scores of TTvSense for VSD  
with and without sentence splitting

TTvSense, which is based on fastSense, has several parameters that must be learned based on the training data. This process of fitting model parameters to existing data is called *model training*. Another class of parameters, called hyperparameters, cannot be learned directly from the training process. Hyperparameters are variables that control the training process itself. They must be set beforehand and are configuration variables of the training process that are kept constant during training. They define higher-level concepts for the model, such as complexity, convergence rate, or penalty (Bergstra and Bengio 2012). We perform hyperparameter optimization to find optimal hyperparameter configurations for TTvSense on TGVCorp that maximize the prediction accuracy. For this task, we use TPE (Bergstra and Bengio 2012) implemented by hyperopt (Bergstra *et al.* 2013). Table 18 shows the parameter space of hyperparameter optimization. Figure 5 shows the results of each trial during the optimization process. The difference between the best and worst performer is 23%. This shows that optimizing the hyperparameters can be crucial.

Table 18: Parameter space of TTVSense used in our experiments. The column *Possible Values* describes the range of values of the parameters. The parameter setting with the best value is highlighted in bold

Parameter	Possible Values
epoch	[5,10,..., <b>40</b> ,...,250]
wordNgram	[1,2,..., <b>10</b> ]
minCount	[1,2,3]
learning rate	[0.1,..., <b>0.2</b> ,...,1]
loss	[ <b>softmax</b> ,hs,ns]
pretrainedVectors	[ <b>true</b> ,false]

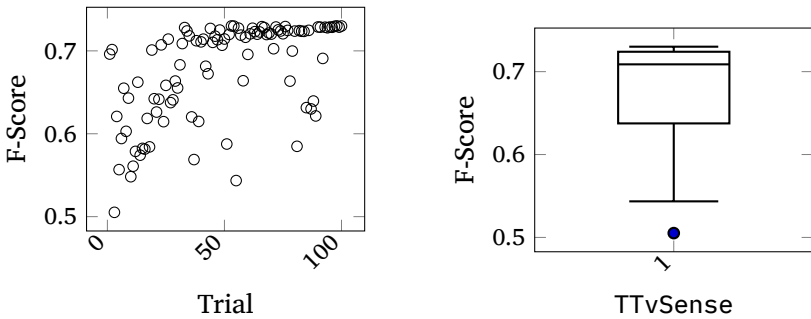


Figure 5: The figure shows the results of optimizing TTVSense on TGVCorp by means of TPE. The scatter plot on the left side shows the results of each trial. The boxplot shows in which area the results are located and how they are distributed over this area. The difference between the best and the worst performing setting is 23%

4

CONCLUSION

In this paper, we have (further) developed an essentially three-part pipeline for VSD in German (1) starting from the constraint-based selection of a part of a sense inventory (i.e. GermaNet) via (2) the annotation of a sense-disambiguated corpus (TGVCorp) to (3) a classifier (TTvSense) trained on it. We also optimized our classifier in three ways: (A) in terms of compressing the selected sense inventory, (B) in terms of obtaining additional training sentences, and

(C) – quasi-traditionally – in terms of hyperparameter optimization. (A) was used to obtain training examples by transfer for senses for which there are not enough annotations in the training corpus. (B) was used to extend our training corpus by generating new sentences. While (A) directly addresses the data bottleneck problem in WSD (Navigli 2009), this does not necessarily apply to (B). The reason for this is probably that sentence generation as we have implemented it only intensifies existing imbalances in the training data (virtually by interpolating along sufficiently confirmed sentence patterns): sentence generation based on our implementation is not creative enough, so to speak. Another outcome of our work is that we disambiguated the occurrences of 1,560 verbs from GermaNet in a corpus based on TüBa-D/Z (see Table 5). As a result, we currently have the largest corpus-based sense-disambiguated set of verbs, for which we simultaneously provide a classifier that outperforms the BERT-based EWISER system in German. This is worth highlighting in two respects: on the one hand, we show a potential for energy saving by relying on a simpler ML architecture to support green NLP (cf. Bender *et al.* 2021). On the other hand, we extend the list of approaches that do not rely on large transformer-based architectures, but instead on simpler resources for solving NLP tasks with comparable quality (cf. Henlein and Mehler 2022, for similar findings).

So far, so traditional our approach. But what about resources that have access to large portions of the web to train the largest possible language models currently available? Don't these methods make NLP efforts like the one shown here seem anachronistic by potentially leveraging access to every online dictionary, every online text that can be linked to it, and every NLP resource that can be used to enhance the database? More precisely, why not just use a large language model such as ChatGPT (OpenAI 2023) as a ready-made tool for NLP including VSD? Why all the effort and tiny technical details when it is so much easier with a tool that seems to have direct access to an all-encompassing resource suitable for almost any NLP task? And indeed, ChatGPT is apparently a ready-made tool also for German VSD. See Figure 6 (left) for a chat in which we embedded Wiktionary's sense inventory for the verb *abtragen* in the sense of *abbauen* 'remove' into a question to ChatGPT that

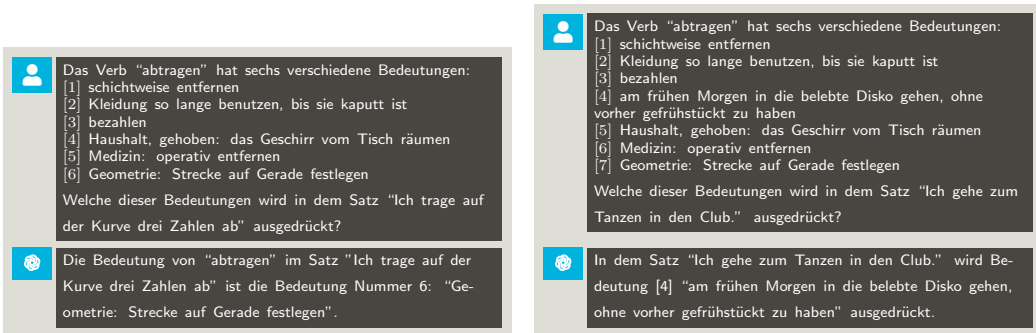


Figure 6: VSD with ChatGPT 3.5 using the Wiktionary entry for the verb *abtragen* 'to dismantle'. We have added an additional fake sense on the right (namely sense [4]), demonstrating that ChatGPT hallucinates (download Wiktionary data/ChatGPT: January 27, 2023 – graphically customized)

answers correctly. One might now assume, and the current discussion suggests, that ChatGPT solves many of the good old computational linguistic tasks for which a large community of researchers has developed so much in the past. Indeed, this could be a realistic scenario if ChatGPT were completely open so that one could reconstruct its responses algorithmically, extend the underlying algorithm as needed, or modify its training resources to adapt it for further research. This apparent gap leaves a third scenario: using ChatGPT to generate training corpora with which to train simple classifiers such as the one presented here, to obtain systems that are at least algorithmically open and that the scientific community can independently develop and adapt for its purposes. Research based on machine reading comprehension (Wang *et al.* 2022) aims in such a direction: it could help public research benefit from the increasingly powerful language models that have themselves benefited from decades of work by a wide range of researchers. In terms of lexical resources, such an open NLP would follow the third and the fifth of the seven theses of Storrer (2001, p. 63, 65) on digital dictionaries: these resources should be transparent (as well as reconstructable or reproducible) and comprehensible for their users, but also expandable according to their own scientific goals. Along this line of thinking, we could add an eighth thesis, namely that NLP resources should be algorithmically controllable and algorithmically extensible by their users. Last

but not least, we return to Figure 6: on the right side, one can see almostw the same chat, except that we have inserted a “nonsense” sense (number 4), which is “correctly” recognized by ChatGPT for an appropriately phrased example sentence without any occurrence of the verb *abtragen*. Such a scenario – which exposes certain capabilities of ChatGPT as an illusion in the minds of its users – brings us back to Section 1 and the question of sense identification: If we believe in the existence, identifiability, and separability of, e.g., word senses (unlike, e.g., Kilgarriff 1997), this task seems to remain a human one, unless we trust the validity of cluster algorithms (or related approaches) operating on, say, vector representations of words (see Schütze 1998 for a seminal work in this regard) to solve this task on a human level. According to this reading, interpretation – and thus, for instance, the determination of relevant word senses – remains a task that cannot yet be automated given the state-of-the-art in ML, not even by resorting to the huge amount of digitized data.

## APPENDICES

### TABLE OF MERGED SENSES

A

The following table shows merged senses, where merging follows one of these decision criteria (C.):

- Senses not distinguishable
- Circular Senses
- Senses/distinctions are missing
- Obsolete or dialectical meanings
- Metaphor

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
78225	76100	ablehnen	■	74690	74898	aufregen	■
79173	78279	ablehnen	■	144916	74898	aufregen	■
78263	78279	ablehnen	■	82315	77888	aufspüren	■
83482	83480	abschließen	■	80824	80818	aufstellen	■
144567	144566	abspielen	■	83259	78652	aufstellen	■
75468	75463	abstimmen	■	82739	81866	auftauchen	■
77711	74980	agieren	■	77554	81866	auftauchen	■
75668	74980	agieren	■	85538	75835	aufteilen	■
79573	74040	anbieten	■	75671	75667	auftreten	■
75755	74040	anbieten	■	83814	82740	auftreten	■
76330	83407	anfangen	■	82725	74394	aufweisen	■
83272	78924	anführen	■	84888	84886	ausbauen	■
79800	79740	angehen	■	84887	84886	ausbauen	■
79517	78181	anlocken	■	83156	78555	ausdenken	■
76490	74114	annehmen	■	77474	74521	aushalten	■
75163	74114	annehmen	■	77462	74521	aushalten	■
77336	77249	annehmen	■	83426	83190	auslösen	■
79535	78077	anordnen	■	145113	76111	ausschalten	■
83780	75422	anpassen	■	78829	78613	aussprechen	■
82446	82402	ansehen	■	145187	84768	austauschen	■
82445	82402	ansehen	■	145195	83519	ausweichen	■
75659	144803	ansiedeln	■	73494	73491	auszeichnen	■
80564	76263	anwenden	■	82930	82896	bauen	■
77735	76263	anwenden	■	77382	79034	beanspruchen	■
144832	75543	anzeigen	■	74672	74678	bedauern	■
77955	77709	arbeiten	■	82700	80406	bedecken	■
79738	79207	attackieren	■	74853	73640	beeindrucken	■
75850	83145	aufbauen	■	84840	78080	beeinflussen	■
78434	85400	aufdecken	■	84870	79663	beeinträchtigen	■
79554	76194	auferlegen	■	145236	80003	befestigen	■
83470	79874	aufgeben	■	76443	76256	befriedigen	■
83497	85392	aufheben	■	82286	77712	begegnen	■
83504	73727	aufhören	■	82320	75176	begegnen	■
77580	77882	aufklären	■	83406	145239	beginnen	■
78832	77882	aufklären	■	81169	75945	begleiten	■
82438	77430	aufpassen	■	109526	79013	begründen	■



*On German verb sense disambiguation*

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
79021	78337	beharren	■	77378	85724	brauchen	■
79766	77478	behaupten	■	85727	85724	brauchen	■
109404	79094	bekräftigen	■	84250	83725	brechen	■
145263	79803	bekämpfen	■	76300	83725	brechen	■
76219	73964	belohnen	■	81248	73921	bringen	■
77420	75553	bemühen	■	78032	73765	charakterisieren	■
78041	75368	benennen	■	78975	78552	darlegen	■
85957	76270	benutzen	■	73766	73304	darstellen	■
77750	78343	berücksichtigen	■	78976	78551	darstellen	■
74239	75567	beschaffen	■	78954	78551	darstellen	■
76509	77950	beschäftigen	■	109332	78593	demonstrieren	■
109437	79935	besetzen	■	77708	77789	denken	■
109435	79935	besetzen	■	83258	78596	dokumentieren	■
75566	75031	besorgen	■	82808	82055	drehen	■
145311	78029	bestimmen	■	81914	82055	drehen	■
109454	75372	bestimmen	■	83349	79622	drucken	■
78328	78324	bestätigen	■	81188	80691	drängen	■
79082	78324	bestätigen	■	75872	75023	durchführen	■
77483	75262	besuchen	■	75866	75023	durchführen	■
141358	76528	betreffen	■	79887	76367	durchsetzen	■
75802	75324	betreiben	■	76240	73457	eignen	■
80757	80753	bewegen	■	78345	73551	einbeziehen	■
78441	82734	beweisen	■	77752	73551	einbeziehen	■
78598	82734	beweisen	■	77963	75164	eingehen	■
109317	73988	bezahlen	■	77373	77361	einrichten	■
109316	73988	bezahlen	■	85175	74094	einräumen	■
77734	79049	beziehen	■	76493	76492	einsetzen	■
76533	79049	beziehen	■	77362	75462	einstellen	■
74039	75746	bieten	■	144378	74209	empfangen	■
83873	75746	bieten	■	82487	74485	empfinden	■
75779	75746	bieten	■	83548	83535	enden	■
79585	77993	billigen	■	82306	77588	entdecken	■
79164	75057	binden	■	83174	78984	entfalten	■
82299	82303	blicken	■	78044	76437	entscheiden	■
85323	76113	blockieren	■	76222	73963	entschädigen	■
85315	76113	blockieren	■	76442	73437	entsprechen	■

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
83158	78543	entwerfen	■	78740	75095	festlegen	■
83036	78535	entwickeln	■	82261	77584	feststellen	■
84008	83834	entwickeln	■	77892	77584	feststellen	■
83882	83834	entwickeln	■	82307	77891	finden	■
109986	74318	erarbeiten	■	81546	81620	fliegen	■
74571	74547	erfreuen	■	141265	81350	fliegen	■
73413	76454	erfüllen	■	79030	77376	fordern	■
78581	73745	ergeben	■	112657	78321	freigeben	■
74434	73745	ergeben	■	74620	74602	fürchten	■
84937	77818	ergänzen	■	75118	73801	geben	■
83883	78308	erheben	■	81724	81356	gehen	■
74724	77109	erholen	■	130725	73519	gehen	■
84039	84038	erhöhen	■	73387	73375	geschehen	■
82264	82262	erkennen	■	78313	76090	gestatten	■
78970	78895	erklären	■	78313	76090	gestatten	■
89997	74211	erlangen	■	77245	77229	glauben	■
76088	78311	erlauben	■	82690	82239	glänzen	■
77545	75260	erleben	■	78194	73600	halten	■
77541	75260	erleben	■	77745	73600	halten	■
79714	74515	erleiden	■	77593	73600	halten	■
74657	74515	erleiden	■	77652	76286	halten	■
77886	82321	ermitteln	■	74370	73671	halten	■
82764	76087	ermöglichen	■	73856	73815	handeln	■
79193	79923	erobern	■	83800	77800	heben	■
110251	78567	erschließen	■	83793	84749	heilen	■
100797	74609	erschrecken	■	82323	77583	herausfinden	■
77454	74518	ertragen	■	78781	78775	hervorheben	■
77331	77396	erwarten	■	79668	76127	hindern	■
74237	74322	erwerben	■	75265	75216	hingehen	■
78960	78959	erzählen	■	77991	74519	hinnehmen	■
83450	75849	eröffnen	■	82728	78787	hinweisen	■
144397	83148	etablieren	■	82450	82447	hören	■
81239	81559	fahren	■	77481	82447	hören	■
81634	81559	fahren	■	74870	78174	inspirieren	■
87060	73571	fehlen	■	77244	77241	kennen	■
87224	84801	festigen	■	77242	77241	kennen	■

*On German verb sense disambiguation*

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
74728	82603	klagen	■	73793	73792	neigen	■
80318	80310	klopfen	■	79449	76412	nennen	■
78529	78522	klären	■	74900	74688	nerven	■
84083	73789	kommen	■	78357	75851	organisieren	■
77713	79643	konfrontieren	■	75569	74255	organisieren	■
78129	75814	kontrollieren	■	141981	80361	packen	■
83243	85706	kopieren	■	112508	112507	probieren	■
82863	85706	kopieren	■	82766	78517	produzieren	■
141069	79789	kämpfen	■	142056	75742	promovieren	■
81843	81834	landen	■	142072	75735	qualifizieren	■
81449	81357	laufen	■	86970	85872	rauchen	■
83806	73401	laufen	■	110711	75589	regeln	■
109367	76423	lauten	■	141611	82907	rekonstruieren	■
73265	76674	leben	■	85174	75822	räumen	■
83944	74723	legen	■	82749	78518	schaffen	■
86971	75707	lehren	■	82781	78518	schaffen	■
79287	77523	lesen	■	129735	79300	schimpfen	■
82677	82207	leuchten	■	85814	129775	schmecken	■
79516	74501	locken	■	87037	74801	schreien	■
78179	74501	locken	■	141668	84827	schwächen	■
140156	77196	locken	■	79748	76018	schützen	■
78509	76298	lösen	■	74386	74371	sparen	■
78426	76298	lösen	■	83017	74363	speichern	■
77579	76298	lösen	■	79286	78950	sprechen	■
83092	83110	malen	■	81463	80765	springen	■
86797	86794	melden	■	82488	74489	spüren	■
86796	86794	melden	■	80952	80958	stammen	■
110714	80694	mischen	■	80957	80958	stammen	■
77600	82281	mitbekommen	■	83441	75871	starten	■
75241	75250	mitmachen	■	130045	74497	staunen	■
74249	81171	mitnehmen	■	79999	80440	stecken	■
140604	80058	montieren	■	80446	80440	stecken	■
74626	73584	mögen	■	89378	89380	stecken	■
77590	78574	nehmen	■	84903	77806	steigern	■
85914	74109	nehmen	■	80844	80813	stellen	■
80339	74109	nehmen	■	82666	76837	stinken	■

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
81861	83502	stoppen	■	84659	79919	vernichten	■
81201	81093	stoßen	■	84262	79919	vernichten	■
82679	82208	strahlen	■	131539	78078	verordnen	■
145181	83179	strahlen	■	112413	75223	verpassen	■
141822	79772	streiten	■	112409	75223	verpassen	■
83764	84804	stärken	■	79171	75099	verpflichten	■
89400	79649	stören	■	78744	78812	verraten	■
73751	75953	stützen	■	81224	81074	verschieben	■
81986	75683	tanzen	■	75762	79159	versprechen	■
75197	77276	trauen	■	89447	78850	verständigen	■
75273	75175	treffen	■	79792	79744	verteidigen	■
89423	89422	treten	■	76011	79011	verteidigen	■
130357	82360	umsehen	■	78361	85576	verteilen	■
83973	75159	unterbringen	■	132277	75434	vertragen	■
130381	75863	unternehmen	■	82963	76271	verwenden	■
86282	79915	unterwerfen	■	77400	79042	vorbehalten	■
78656	76196	urteilen	■	132404	79808	vordringen	■
78729	75108	verabschieden	■	112510	82326	vorfinden	■
130400	85386	verbergen	■	78653	75694	vorgeben	■
84688	83789	verbessern	■	77366	77365	vorsehen	■
82970	85720	verbrauchen	■	78570	77596	vorstellen	■
110875	74215	verbuchen	■	76414	76413	vorstellen	■
81361	77414	verfolgen	■	132715	78967	vortragen	■
79560	74337	verfügen	■	82721	73940	vorweisen	■
78031	74337	verfügen	■	109707	109708	wachen	■
74405	74337	verfügen	■	84007	76735	wachsen	■
130457	75012	vergewaltigen	■	83859	84024	wachsen	■
73296	73645	verhalten	■	80590	84998	wachsen	■
130471	78674	verhandeln	■	77275	75194	wagen	■
78804	75070	verheiraten	■	83556	73391	wandeln	■
84852	84067	verkürzen	■	78918	78913	warnen	■
75925	77318	verlangen	■	89494	73656	warten	■
83938	74423	verlieren	■	76055	73656	warten	■
84003	84923	verlängern	■	73824	73823	wechseln	■
112505	75571	vermitteln	■	89501	84143	wechseln	■
111004	76022	vernachlässigen	■	85060	74157	wegnehmen	■

*On German verb sense disambiguation*

LexIds	maps to	lemma	C.	LexIds	maps to	lemma	C.
132876	82251	wehen	■	79069	78227	zugeben	■
112234	79746	wehren	■	78315	76091	zulassen	■
133237	79595	weiterleiten	■	139606	78532	zurückführen	■
133293	133286	wenden	■	74848	73307	zusammenhängen	■
109333	79199	werben	■	75845	74281	zusammenstellen	■
84147	77510	wiederholen	■	78533	78004	zuschreiben	■
113289	82738	wiederspiegeln	■	139871	77996	zustimmen	■
73643	73637	wirken	■	84160	78231	ändern	■
83180	73637	wirken	■	78608	78742	äußern	■
73329	73312	wohnen	■	83970	85366	öffnen	■
74008	73967	zahlen	■	83965	85376	öffnen	■
89629	83077	zeichnen	■	73739	73831	überlassen	■
83101	83077	zeichnen	■	74111	74110	übernehmen	■
73628	78592	zeigen	■	130392	73677	übersehen	■
113100	78428	zerlegen	■	82436	76079	überwachen	■
81203	81075	ziehen	■	139979	76299	überwinden	■

## B

## RESOURCE VERSIONS

This appendix lists the details on the corpora we used, in particular the version or date accessed.

1. **BabelNet** – Version 4.0.1
2. **Bundestag Corpus** – Full texts of the plenary minutes and printed papers of the German Bundestag from the 1st to the 18th legislative period (1949–2017)
3. **COW** – decow16ax (DE stands for German, COW for “CORpus from the Web”, 16 for 2016 (major technology version), A for the first release built using 2016 technology. The following X indicates that the corpus is a sentence shuffle)
4. **COW16b** – decow16bx (DE stands for German, COW for “CORpus from the Web”, 16 for 2016 (major technology version), B for the second release built using 2016 technology. The following X indicates that the corpus is a sentence shuffle)
5. **DeReKo** – We did not have access to this corpus directly, due to licensing issues. Instead, the *Institut für Deutsche Sprache* (IDS) kindly sent us a summary of frequency, lemma and POS information for tokens occurring in a section (DeReKo-2020-I subcorpus) of the full corpus
6. **deWaC** – <https://wacky.sslmit.unibo.it> (Baroni et al. 2009)
7. **DTA** – *Deutsches Textarchiv*. Core and supplementary texts, version released on July 21, 2017
8. **Duden** – *Deutsches Universalwörterbuch* 2003; for exemplification we additionally consulted the Duden online version (download: 2024-02-14)
9. **EU Bookshop** – Release v2 (Tiedemann 2012)
10. **E-VALBU** – final version
11. **Gutenberg** – Edition 13
12. **GermaNet** – Version 14
13. **GVSD** – *The German Verb Subcategorisation Database*. Accessed on February 15, 2021
14. **Leipziger Wortschatz** – volumes 1995–1997 (Goldhahn et al. 2012)
15. **Textbooks** – A collection of 14 German textbooks on economics, published between 2014 and 2020. The textbooks have been used in the study by Lücking et al. (2021) and are listed in their appendix B
16. **SALSA** – SALSA 2.0
17. **Süddeutsche Zeitung** – 1992–2014
18. **TüBa-D/Z** – Version 10.0
19. **WebCAGe** – Version 3.0
20. **Wikipedia** – German version, accessed on February 3, 2016.
21. **Wiktionary** – German version, accessed on May 1, 2019.
22. **Die ZEIT** – 1946–2007

## REFERENCES

- Giuseppe ABRAMI, Mevlüt BAGCI, Leon HAMMERLA, and Alexander MEHLER (2022), German Parliamentary Corpus (GerParCor), in *Proceedings of the Language Resources and Evaluation Conference (LREC 2022)*, pp. 1900–1906, European Language Resources Association, Marseille, France.
- Jeff ALSTOTT, Ed BULLMORE, and Dietmar PLENZ (2014), powerlaw: a Python package for analysis of heavy-tailed distributions, *PLoS ONE*, 9(4):e95816, doi:10.1371/journal.pone.0095816.
- Nicholas ASHER, Márta ABRUSÁN, and Tim VAN DE CRUYS (2017), Types, meanings and co-composition in lexical semantics, in Stergios CHATZIKYRIAKIDIS and Zhaohui LUO, editors, *Modern perspectives in type-theoretical semantics*, number 98 in Studies in Linguistics and Philosophy, pp. 135–161, Springer International Publishing AG, Cham, Switzerland, doi:10.1007/978-3-319-50422-3\_6.
- Collin F. BAKER, Charles J. FILLMORE, and John B. LOWE (1998), The Berkeley FrameNet project, in *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING-ACL '98, August 10–14, 1998, Université de Montréal, Montréal, Quebec, Canada. Proceedings of the Conference.*, pp. 86–90, <http://aclweb.org/anthology/P/P98/P98-1013.pdf>.
- Marco BARONI, Silvia BERNARDINI, Adriano FERRARESI, and Eros ZANCHETTA (2009), The WaCky wide web: A collection of very large linguistically processed web-crawled corpora, *Language Resources & Evaluation*, 43:209–226, doi:10.1007/s10579-009-9081-4.
- Emily M. BENDER, Timnit GEBRU, Angelina MCMILLAN-MAJOR, and Shmargaret SHMITCHELL (2021), On the dangers of stochastic parrots: Can language models be too big?, in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, p. 610–623, Association for Computing Machinery, New York, NY, USA, ISBN 9781450383097, doi:10.1145/3442188.3445922.
- James BERGSTRA and Yoshua BENGIO (2012), Random search for hyper-parameter optimization, *Journal of Machine Learning Research*, 13:281–305, <http://dl.acm.org/citation.cfm?id=2188395>.
- James BERGSTRA, Daniel YAMINS, and David D. COX (2013), Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 115–123, <http://proceedings.mlr.press/v28/bergstra13.html>.

Michele BEVILACQUA and Roberto NAVIGLI (2020), Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2854–2864, Association for Computational Linguistics, Online, doi:10.18653/v1/2020.acl-main.255, <https://aclanthology.org/2020.acl-main.255>.

Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN, and Tomáš MIKOLOV (2016), Enriching word vectors with subword information, *CoRR*, abs/1607.04606, <http://arxiv.org/abs/1607.04606>.

Claire BONIAL, Julia BONN, Kathryn CONGER, Jena HWANG, Martha PALMER, and Nicholas REESEM (2015), English PropBank annotation guidelines, Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder, <http://propbank.github.io/>.

Zdravko BOTEV and Ad RIDDER (2017), *Variance reduction*, pp. 1–6, American Cancer Society, ISBN 9781118445112, doi:10.1002/9781118445112.stat07975, <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat07975>.

Sabine BRANTS, Stefanie DIPPER, Peter EISENBERG, Silvia HANSEN, Esther KÖNIG, Wolfgang LEZIUS, Christian ROHRER, George SMITH, and Hans USZKOREIT (2004), TIGER: Linguistic interpretation of a German corpus, *Journal of Language and Computation*, 2:597–620.

Aljoscha BURCHARDT, Katrin ERK, Anette FRANK, Andrea KOWALSKI, Sebastian PADÓ, and Manfred PINKAL (2006), The SALSA corpus: A German corpus resource for lexical semantics, in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, European Language Resources Association (ELRA), Genoa, Italy, [http://www.lrec-conf.org/proceedings/lrec2006/pdf/339\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/339_pdf.pdf).

Gennaro CHIERCHIA and Sally MCCONNELL-GINET (2000), *Meaning and grammar – an introduction to semantics*, MIT Press, Cambridge, 2 edition.

Aaron CLAUSET, Cosma Rohilla SHALIZI, and Mark E. J. NEWMAN (2009), Power-law distributions in empirical data, *SIAM Review*, 51(4):661–703, doi:10.1137/070710111, Society of Industrial and Applied Mathematics.

Robin COOPER (2011), Copredication, quantification and frames, in Sylvain POGODALLA and Jean-Philippe PROST, editors, *Logical aspects of computational linguistics*, number 6736 in Lecture Notes in Computer Science, pp. 64–79, Springer, Berlin and Heidelberg, doi:10.1007/978-3-642-22221-4\_5.

D. Alan CRUSE (2000), *Meaning in language*, Oxford University Press, New York.



Gábor CsÁRDI and Tamás NEPU SZ (2006), The igraph software package for complex network research, *InterJournal*, Complex Systems:1695, <https://igraph.org>.

Benjamin DAVID, Sylvia SPRINGORUM, and Sabine SCHULTE IM WALDE (2014), German perception verbs: Automatic classification of prototypical and multiple non-literal meanings, in *Proceedings of the 12th Konvens 2014*.

Henriette DE SWART (2011), Mismatches and coercion, in Claudia MAIENBORN, Klaus VON HEUSINGER, and Paul PORTNER, editors, *Semantics: An international handbook of natural language meaning*, volume 1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, chapter 25, pp. 574–597, De Gruyter Mouton, doi:10.1515/9783110226614.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2018), BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805*.

Andrea DI FABIO, Simone CONIA, and Roberto NAVIGLI (2019), VerbAtlas: A novel large-scale verbal semantic resource and its application to semantic role labeling, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 627–637, Association for Computational Linguistics, Hong Kong, China, doi:10.18653/v1/D19-1058, <https://www.aclweb.org/anthology/D19-1058>.

Stefanie DIPPER, Hannah KERMES, Esther KÖNIG-BAUMER, Wolfgang LEZIUS, Frank H. MÜLLER, and Tylman ULE (2002), DEREKO – (DEutsches REferenzKOrpus) German Reference Corpus. Final report (Part I), Technical report, IMS: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, SfS: Seminar für Sprachwissenschaft, Universität Tübingen.

Konrad DUDEN, Dieter BERGER, and Werner SCHOLZE (1980), *Duden*, volume 2, Bibliographisches Institut.

Veena D. DWIVEDI (2013), Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing, *PLoS ONE*, 8(11):e81461, doi:10.1371/journal.pone.0081461.

Philip EDMONDS and Scott COTTON (2001), SENSEVAL-2: overview, in *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL@ACL 2001, Toulouse, France, July 5-6, 2001*, pp. 1–5, <https://aclanthology.info/papers/S01-1001/s01-1001>.

Gertrud FAASS and Kerstin ECKART (2013), SdeWaC – a corpus of parsable sentences from the web, in Iryna GUREVYCH, Chris BIEMANN, and Torsten ZESCH, editors, *Language processing and knowledge in the web*, pp. 61–68, Springer Berlin Heidelberg, Berlin, Heidelberg, ISBN 978-3-642-40722-2.

Christiane FELLBAUM, editor (1998), *WordNet: An electronic lexical database*, MIT Press, Cambridge.

Christiane FELLBAUM and George A. MILLER (1998), *Lexical chains as representations of context for the detection and correction of malapropisms*, pp. 305–332, MITP, ISBN 9780262272551, <https://ieeexplore.ieee.org/document/6287673>.

Christiane FELLBAUM, Martha PALMER, Hoa Trang DANG, Lauren DELFS, and Susanne WOLF (2001), Manual and automatic semantic annotation with WordNet, in *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*, pp. 1–8, Carnegie Mellon University Pittsburgh, PA.

Charles J. FILLMORE and Colin BAKER (2010), A frames approach to semantic analysis, in Bernd HEINE and Heiko NARROG, editors, *The Oxford Handbook of Linguistic Analysis*, pp. 313–340, Oxford University Press, Oxford.

William A. GALE, Kenneth W. CHURCH, and David YAROWSKY (1992), One sense per discourse, in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, <https://www.aclweb.org/anthology/H92-1045>.

Spandana GELLA, Desmond ELLIOTT, and Frank KELLER (2019), Cross-lingual visual verb sense disambiguation, in *Proceedings of NAACL-HLT 2019*, pp. 1998–2004.

Brendan S. GILLON (1990), Ambiguity, generality, and indeterminacy: Tests and definitions, *Synthese*, 85(3):391–416.

Dirk GOLDHAHN, Thomas ECKART, and Uwe QUASTHOFF (2012), Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages, in Nicoletta CALZOLARI, Khalid CHOUKRI, Thierry DECLERCK, Mehmet Ugur DOGAN, Bente MAEGAARD, Joseph MARIANI, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, ISBN 978-2-9517408-7-7.

Birgit HAMP and Helmut FELDWEG (1997), GermaNet – a lexical-semantic net for German, in *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pp. 9–15.

Wahed HEMATI (2020), *TextImager-VSD: Large scale verb sense disambiguation and named entity recognition in the context of TextImager*, Ph.D. thesis, Goethe-University Frankfurt, <http://publikationen.uni-frankfurt.de/frontdoor/index/index/docId/56089>.

Wahed HEMATI, Tolga USLU, and Alexander MEHLER (2016), TextImager: A distributed UIMA-based system for NLP, in *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, December 11–16, 2016, Osaka, Japan*, pp. 59–63, <https://www.aclweb.org/anthology/C16-2013/>.

Alexander HENLEIN and Alexander MEHLER (2022), What do toothbrushes do in the kitchen? How transformers think our world is structured, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5791–5807, Association for Computational Linguistics, Seattle, United States, doi:10.18653/v1/2022.naacl-main.425, <https://aclanthology.org/2022.naacl-main.425>.

Verena HENRICH (2015), *Word sense disambiguation with GermaNet*, Ph.D. thesis, Universität Tübingen, doi:10.15496/publikation-4706, <https://publikationen.uni-tuebingen.de/xmlui/handle/10900/63284>.

Verena HENRICH and Erhard HINRICHS (2012), A comparative evaluation of word sense disambiguation algorithms for German, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 576–583, European Language Resources Association (ELRA), Istanbul, Turkey, [http://www.lrec-conf.org/proceedings/lrec2012/pdf/164\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2012/pdf/164_Paper.pdf).

Verena HENRICH and Erhard W. HINRICHS (2013), Extending the TüBa-D/Z Treebank with GermaNet sense annotation, in *Language processing and knowledge in the web – 25th International Conference, GSCL 2013, Darmstadt, Germany, September 25–27, 2013. Proceedings*, pp. 89–96, doi:10.1007/978-3-642-40722-2\_9.

Verena HENRICH, Erhard W. HINRICHS, and Tatiana VODOLAZOVA (2011), Aligning GermaNet senses with Wiktionary sense definitions, in *Human Language Technology Challenges for Computer Science and Linguistics – 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25–27, 2011, Revised Selected Papers*, pp. 329–342, doi:10.1007/978-3-319-08958-4\_27.

Verena HENRICH, Erhard W. HINRICHS, and Tatiana VODOLAZOVA (2012), Webcage – A web-harvested corpus annotated with GermaNet senses, in *EACL 2012, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, April 23–27, 2012*, pp. 387–396, <http://aclweb.org/anthology/E/E12/E12-1039.pdf>.

Ray JACKENDOFF (1983), *Semantics and cognition*, MIT Press, Cambridge, MA.

Armand JOULIN, Edouard GRAVE, Piotr BOJANOWSKI, and Tomas MIKOLOV (2017), Bag of tricks for efficient text classification, in *Proceedings of the 15th Conference of the EACL: Volume 2, Short Papers*, pp. 427–431, Association for Computational Linguistics, Valencia, Spain, <https://www.aclweb.org/anthology/E17-2068>.

David KAPLAN (1989), Demonstratives, in Joseph ALMOG, John PERRY, and Howard WETTSTEIN, editors, *Themes from Kaplan*, pp. 481–563, Oxford University Press, New York and Oxford.

Rudi KELLER (1990), *Sprachwandel: von der unsichtbaren Hand in der Sprache*, Francke, Tübingen.

Christopher KENNEDY (2011), Ambiguity and vagueness: An overview, in Claudia MAIENBORN, Klaus VON HEUSINGER, and Paul PORTNER, editors, *Semantics: An international handbook of natural language meaning*, volume 1 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, chapter 23, pp. 507–535, De Gruyter Mouton, doi:10.1515/9783110226614.

Adam KILGARRIFF (1997), “I don’t believe in word senses”, *Computers and the Humanities*, 31(2):91–113, doi:10.1023/A:1000583911091.

Adam KILGARRIFF (1998), Gold standard datasets for evaluating word sense disambiguation programs, *Computer Speech & Language*, 12(4):453–472, doi:10.1006/csla.1998.0108.

Paul R. KROEGER (2019), *Analyzing meaning*, number 5 in Textbooks in Language Sciences, Language Science Press, Berlin, second corrected and slightly revised edition.

Jacqueline KUBCZAK (2009), Hier wird Ihnen geholfen! E-VALBU – Das elektronische Valenzwörterbuch deutscher Verben, *Sprachreport*, 4:17–23.

Claudia KUNZE and Lothar LEMNITZER (2002), GermaNet – representation, visualization, application, in M. RODRIGUEZ GONZÁLEZ and C. PAZ SUÁREZ ARAUJO, editors, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1485–1491, European Language Resources Association, Paris.

Els LEFEVER and Véronique HOSTE (2010), SemEval-2010 task 3: Cross-lingual word sense disambiguation, in *5th International Workshop on Semantic Evaluation (SemEval 2010)*, pp. 15–20, Association for Computational Linguistics (ACL).

Els LEFEVER and Véronique HOSTE (2013), SemEval-2013 task 10: Cross-lingual word sense disambiguation, in *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 158–166.

Beth LEVIN (1993), *English verb classes and alternations: A preliminary investigation*, University of Chicago Press.

Andy LÜCKING, Sebastian BRÜCKNER, Giuseppe ABRAMI, Tolga USLU, and Alexander MEHLER (2021), Computational linguistic assessment of textbooks and online texts by means of threshold concepts in economics, *Frontiers in Education*, 5:578475, doi:10.3389/educ.2020.578475, <https://www.frontiersin.org/articles/10.3389/educ.2020.578475/>.

Fuli LUO, Tianyu LIU, Qiaolin XIA, Baobao CHANG, and Zhifang SUI (2018), Incorporating glosses into neural word sense disambiguation, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, Volume 1: Long Papers*,

pp. 2473–2482,

<https://aclanthology.info/papers/P18-1230/p18-1230>.

John LYONS (1977), *Semantics*, volume 1, Cambridge University Press, London.

Alexander MEHLER, Rüdiger GLEIM, Wahed HEMATI, and Tolga USLU (2018), Skalenfreie online soziale Lexika am Beispiel von Wiktionary, in Stefan ENGELBERG, Henning LOBIN, Kathrin STEYER, and Sascha WOLFER, editors, *Proceedings of 53rd Annual Conference of the Institut für Deutsche Sprache (IDS), March 14–16, Mannheim, Germany*, pp. 269–291, De Gruyter, Berlin.

Oren MELAMUD, Jacob GOLDBERGER, and Ido DAGAN (2016), context2vec: Learning generic context embedding with bidirectional LSTM, in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11–12, 2016*, pp. 51–61, <http://aclweb.org/anthology/K/K16/K16-1006.pdf>.

George A. MILLER (1995), Wordnet: A lexical database for English, *Communications of the ACM*, 38(11):39–41, doi:10.1145/219717.219748, <http://doi.acm.org/10.1145/219717.219748>.

George A. MILLER, Martin CHODOROW, Shari LANDES, Claudia LEACOCK, and Robert G. THOMAS (1994), Using a semantic concordance for sense identification, in *Human Language Technology: Proceedings of a workshop held at Plainsboro, New Jersey, March 8–11, 1994*, <https://aclanthology.org/H94-1046>.

Marc MOENS and Mark STEEDMAN (1988), Temporal ontology and temporal reference, *Computational Linguistics*, 14(2):15–28.

Roberto NAVIGLI (2009), Word sense disambiguation: A survey, *ACM Computing Survey*, 41(2):10:1–10:69, doi:10.1145/1459352.1459355, <https://doi.org/10.1145/1459352.1459355>.

Roberto NAVIGLI, José CAMACHO-COLLADOS, and Alessandro RAGANATO (2017), Word sense disambiguation: A unified evaluation framework and empirical comparison, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3–7, 2017, Volume 1: Long Papers*, pp. 99–110, <https://aclanthology.info/papers/E17-1010/e17-1010>.

Roberto NAVIGLI, David JURGENS, and Daniele VANNELLA (2013), SemEval-2013 task 12: Multilingual word sense disambiguation, in *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14–15, 2013*, pp. 222–231, <http://aclweb.org/anthology/S/S13/S13-2040.pdf>.

Roberto NAVIGLI and Simone Paolo PONZETTO (2012), BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial Intelligence*, 193:217–250, ISSN 0004-3702, doi:10.1016/j.artint.2012.07.001.

- Mark E. J. NEWMAN (2005), Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, 46:323–351.
- Geoffrey NUNBERG (1995), Transfers of meaning, *Journal of Semantics*, 12(2):109–132, doi:10.1093/jos/12.2.109.
- OPENAI (2023), ChatGPT (version 3.5), <https://github.com/openai/gpt-3>.
- Martha PALMER, Hoa Trang DANG, and Christiane FELLBAUM (2007), Making fine-grained and coarse-grained sense distinctions, both manually and automatically, *Natural Language Engineering*, 13(2):137–163, doi:10.1017/S135132490500402X.
- Martha PALMER, Daniel GILDEA, and Nianwen XUE (2010), *Semantic role labeling*, Morgan & Claypool Publishers.
- Simone PAPANDREA, Alessandro RAGANATO, and Claudio Delli BOVI (2017), SupWSD: A flexible toolkit for supervised word sense disambiguation, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9–11, 2017 – System Demonstrations*, pp. 103–108, <https://aclanthology.info/papers/D17-2018/d17-2018>.
- Rebecca J. PASSONNEAU, Collin F. BAKER, Christiane FELLBAUM, and Nancy IDE (2012), The MASC word sense corpus, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, May 23–25, 2012*, pp. 3025–3030, <http://www.lrec-conf.org/proceedings/lrec2012/summaries/589.html>.
- Matthew E. PETERS, Mark NEUMANN, Mohit IYYER, Matt GARDNER, Christopher CLARK, Kenton LEE, and Luke ZETTMAYER (2018), Deep contextualized word representations, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237, <https://aclanthology.info/papers/N18-1202/n18-1202>.
- Mohammad Taher PILEHVAR and Jose CAMACHO-COLLADOS (2021), *Embeddings in natural language processing: Theory and advances in vector representations of meaning*, Morgan & Claypool Publishers.
- Sameer PRADHAN, Edward LOPER, Dmitriy DLIGACH, and Martha PALMER (2007), Semeval-2007 task 17: English lexical sample, SRL and all words, in *Proceedings of the 4th International Workshop on Semantic Evaluations, SemEval@ACL 2007, Prague, Czech Republic, June 23–24, 2007*, pp. 87–92, <http://aclweb.org/anthology/S/S07/S07-1016.pdf>.
- James PUSTEJOVSKY (1991), The generative lexicon, *Computational Linguistics*, 17:409–441.
- James PUSTEJOVSKY (1995), *The generative lexicon*, MIT Press, Cambridge, MA.

Diana RAILEANU, Paul BUITELAAR, Spela VINTAR, and Jörg BAY (2002), Evaluation corpora for sense disambiguation in the medical domain, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), May 29–31, 2002, Las Palmas, Canary Islands, Spain*, <http://www.lrec-conf.org/proceedings/lrec2002/sumarios/166.htm>.

Prajit RAMACHANDRAN, Barret ZOPH, and Quoc V. LE (2017), Searching for activation functions, *CoRR*, abs/1710.05941, <http://arxiv.org/abs/1710.05941>.

Shauli RAVFOGEL, Yanai ELAZAR, Jacob GOLDBERGER, and Yoav GOLDBERG (2020), Unsupervised distillation of syntactic information from contextualized word representations, *arXiv preprint arXiv:2010.05265*.

Burghard B. RIEGER (1989), *Unscharfe Semantik: Die empirische Analyse, quantitative Beschreibung, formale Repräsentation und prozedurale Modellierung vager Wortbedeutungen in Texten*, Peter Lang, Frankfurt a. M.

Burghard B. RIEGER (2001), Computing granular word meanings. A fuzzy linguistic approach in computational semiotics, in Paul WANG, editor, *Computing with words*, pp. 147–208, Wiley, New York.

Sascha ROTHE, Shashi NARAYAN, and Aliaksei SEVERYN (2020), Leveraging pre-trained checkpoints for sequence generation tasks, *Transactions of the Association for Computational Linguistics*, 8:264–280, doi:10.1162/tacl\_a\_00313, <https://aclanthology.org/2020.tacl-1.18>.

Josef RUPPENHOFER, Michael ELLSWORTH, Myriam SCHWARZER-PETRUCK, Christopher R. JOHNSON, and Jan SCHEFFCZYK (2016), FrameNet II: Extended theory and practice, Technical report, International Computer Science Institute.

Jahn-Takeshi SAITO, Joachim WAGNER, Graham KATZ, P. D. Gerson REUTER, Michael B. BURKE, and Sabine REINHARD (2002), Evaluation of GermaNet: Problems using GermaNet for automatic word sense disambiguation, in *LREC Workshop on WordNet Structure and Standardization and How these Affect WordNet Applications and Evaluation*.

Silke SCHEIBLE, Sabine SCHULTE IM WALDE, Marion WELLER, and Max KISSELEW (2013), A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from a web corpus: Tool, guidelines and resource, in *Proceedings of the 8th Web as Corpus Workshop*, pp. 63–72, Lancaster, UK.

Anne SCHILLER, Simone TEUFEL, Christine STÖCKERT, and Christine THIELEN (1999), Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset), Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Karin Kipper SCHULER (2006), *Verbnet: A broad-coverage, comprehensive verb lexicon*, Ph.D. thesis, University of Pennsylvania, <http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf>.

- Helmut SCHUMACHER, editor (1986), *Verben in Feldern. Valenzwörterbuch zur Syntax und Semantik deutscher Verben*, de Gruyter, Berlin and New York.
- Helmut SCHUMACHER, Jacqueline KUBCZAK, Renate SCHMIDT, and Vera DE RUITER (2004), *VALBU – Valenzwörterbuch deutscher Verben*, number 31 in *Studien zur Deutschen Sprache*, Narr, Tübingen.
- Hinrich SCHÜTZE (1998), Automatic word sense discrimination, *Computational Linguistics*, 24(1):97–123.
- Roland SCHÄFER (2015), Processing and querying large web corpora with the COW14 architecture, in *Proceedings of Challenges in the Management of Large Corpora 3 (CMC-3)*, UCREL, IDS, Lancaster, <http://rolandschaefer.net/?p=749>.
- Roland SCHÄFER and Felix BILDHAUER (2012), Building large corpora from the web using a new efficient tool chain, in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 486–493, European Language Resources Association (ELRA), Istanbul, Turkey, ISBN 978-2-9517408-7-7, <http://rolandschaefer.net/?p=70>.
- Benjamin SNYDER and Martha PALMER (2004), The English all-words task, in *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, SENSEVAL@ACL 2004, Barcelona, Spain, July 25-26, 2004*, <https://aclanthology.info/papers/W04-0811/W04-0811>.
- Jan-Philipp SOEHN (2005), Selectional restrictions in HPSG: I'll eat my hat!, in Stefan MÜLLER, editor, *Proceedings of the HPSG05 Conference*, pp. 343–353, Department of Informatics, University of Lisbon, CSLI Publications, Stanford.
- John F. SOWA (2000), *Knowledge representation: Logical, philosophical, and computational foundations*, Brooks/Cole.
- Luc STEELS (2011–12), Modeling the cultural evolution of language, *Physics of Life Reviews*, 8(4):339–356, doi:10.1016/j.plrev.2011.10.014, <http://groups.lis.illinois.edu/amag/langev/paper/steels2011REVIEW.html>.
- Angelika STORRER (2001), Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie, in *Chancen und Perspektiven computer-gestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher*, pp. 53–70, Max Niemeyer Verlag, Tübingen.
- Angelika STORRER (2010), Deutsche Internet-Wörterbücher: Ein Überblick, *Lexicographica*, 27(1):155–164.
- Heike TELLJOHANN, Erhard W. HINRICHS, Sandra KÜBLER, Heike ZINSMEISTER, and Kathrin BECK (2012), *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*, Seminar für Sprachwissenschaft, Wilhelmstr. 19, D-72074 Tübingen, <http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-1201.pdf>.



THE MATHWORKS, INC. (2012), MATLAB and curve fitting toolbox release 2012, Natick, MA.

Jörg TIEDEMANN (2012), Parallel data, tools and interfaces in OPUS, in Nicoletta CALZOLARI, Khalid CHOUKRI, Thierry DECLERCK, Mehmet Ugur DOGAN, Bente MAEGAARD, Joseph MARIANI, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, ISBN 978-2-9517408-7-7.

Tolga USLU, Alexander MEHLER, Daniel BAUMARTZ, Alexander HENLEIN, and Wahed HEMATI (2018), fastsense: An efficient word sense disambiguation classifier, in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Loïc VIAL, Benjamin LECOUEUX, and Didier SCHWAB (2019), Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation, *CoRR*, abs/1905.05677, <http://arxiv.org/abs/1905.05677>.

Piek VOSSEN (1998), Introduction to EuroWordNet, *Computers and the Humanities*, 32(2-3):73–89, doi:10.1023/A:1001175424222.

Nan WANG, Jiwei LI, Yuxian MENG, Xiaofei SUN, Han QIU, Ziyao WANG, Guoyin WANG, and Jun HE (2022), An MRC framework for semantic role labeling, in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2188–2198, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, <https://aclanthology.org/2022.coling-1.191>.

Stephen WECHSLER, Jean-Pierre KOENIG, and Anthony DAVIS (2021), Argument structure and linking, in Stefan MÜLLER, Anne ABEILLÉ, Robert D. BORSLEY, and Jean-Pierre KOENIG, editors, *Head-Driven Phrase Structure Grammar: The handbook*, Language Science Press, Berlin, <https://langsci-press.org/catalog/book/259>, prepublished book chapter.

WIKTIONARY (2019), Free dictionary, <https://www.wiktionary.org/>, accessed: 2019-09-23.

Ian H. WITTEN, Eibe FRANK, and Mark A. HALL (2011), *Data mining: Practical machine learning tools and techniques, 3rd edition*, Morgan Kaufmann, Elsevier, ISBN 9780123748560, <http://www.worldcat.org/oclc/262433473>.

Yonghui WU, Mike SCHUSTER, Zhifeng CHEN, Quoc V. LE, Mohammad NOROUZI, Wolfgang MACHEREY, Maxim KRİKUN, Yuan CAO, Qin GAO, Klaus MACHEREY, Jeff KLINGNER, Apurva SHAH, Melvin JOHNSON, Xiaobing LIU, Lukasz KAISER, Stephan GOUWS, Yoshikiyo KATO, Taku KUDO, Hideto KAZAWA, Keith STEVENS, George KURIAN, Nishant PATIL, Wei WANG, Cliff

YOUNG, Jason SMITH, Jason RIESA, Alex RUDNICK, Oriol VINYALS, Greg CORRADO, Macduff HUGHES, and Jeffrey DEAN (2016), Google’s neural machine translation system: Bridging the gap between human and machine translation, *CoRR*, abs/1609.08144, <http://arxiv.org/abs/1609.08144>.

Yichu ZHOU and Vivek SRIKUMAR (2022), A closer look at how fine-tuning changes BERT, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1046–1061, Association for Computational Linguistics, Dublin, Ireland, doi:10.18653/v1/2022.acl-long.75, <https://aclanthology.org/2022.acl-long.75>.

Arnold M. ZWICKY and Jerrold M. SADOCK (1975), Ambiguity tests and how to fail them, in *Syntax and Semantics volume 4*, pp. 1–36, Academic Press, New York.

*Dominik Mattern*

Text Technology Lab  
Goethe University Frankfurt  
Frankfurt am Main, Germany

*Wahed Hemati*

© 0000-0002-5477-2538  
Shikenso GmbH  
Frankfurt am Main, Germany

*Andy Lücking*

© 0000-0002-5070-2233  
Text Technology Lab  
Goethe University Frankfurt  
Frankfurt am Main, Germany

*Alexander Mehler*

© 0000-0003-2567-7539  
[mehler@em.uni-frankfurt.de](mailto:mehler@em.uni-frankfurt.de)  
Text Technology Lab  
Goethe University Frankfurt  
Frankfurt am Main, Germany  
(Corresponding author)

Dominik Mattern, Wahed Hemati, Andy Lücking, and Alexander Mehler (2024), *On German verb sense disambiguation: A three-part approach based on linking a sense inventory (GermaNet) to a corpus through annotation (TGVCorp) and using the corpus to train a VSD classifier (TTvSense)*, *Journal of Language Modelling*, 12(1):155–212

doi: <https://dx.doi.org/10.15398/jlm.v12i1.356>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

©  <http://creativecommons.org/licenses/by/4.0/>

# QRGS – Question Responses Generation *via* crowdsourcing

*Paweł Łupkowski<sup>1</sup>, Jonathan Ginzburg<sup>2</sup>, Ewelina Chmurska<sup>1</sup>,  
Adrianna Płatosz<sup>1</sup>, Aleksandra Kwiecień<sup>1</sup>, Barbara Adamska<sup>1</sup>,  
and Magdalena Szkalej<sup>1</sup>*

<sup>1</sup> Adam Mickiewicz University

<sup>2</sup> Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle

## ABSTRACT

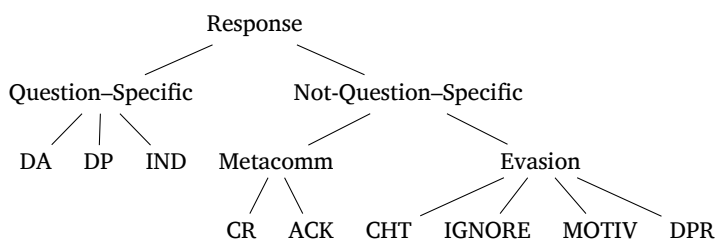
QRGS stands for the Question Responses Generation System. It is an online game-like framework designed for gathering various types of question responses. A QRGS user is asked to read a simple story and impersonate its main character. As the story unfolds the user is confronted with four questions and (s)he is expected to answer these in the way the main character would. In this way, we obtain responses to questions of a desired type. The data gathered *via* QRGS is a useful supplement to the linguistic data already present in language corpora – especially for languages for which such resources are sparse. As such, it opens the possibility for better understanding of the use of questions in natural language dialogues and analysing the response space of such questions. In this paper, we present the main idea of QRGS and the results of five studies (in Polish and in English) that test the framework. Our discussion addresses issues concerning the efficiency and accuracy of the proposed approach. We also discuss the availability of the QRGS and its potential future improvements.

*Keywords:*  
*gamification,*  
*crowdsourcing,*  
*questions,*  
*responses,*  
*language*  
*resources*

This paper describes how certain types of responses to questions (i.e. direct, indirect and evasive ones) may be gathered *via* a relatively simple and easy to use crowdsourcing framework. Question Responses Generation System (QRGS) is designed and implemented with the aim set for providing supplementary data for the study of the response space for questions (Ginzburg *et al.* 2019, 2022).

Ginzburg *et al.* (2019, 2022) present extensive corpus studies of the BNC (Burnard 2007), BEE (Rosé *et al.* 1999), Maptask (Anderson *et al.* 1991) and CornellMovie (Danescu-Niculescu-Mizil and Lee 2011) corpora for English (which include 607, 262, 460, and 911 question/response pairs respectively) and data for Polish using the Spokes corpus (Pezik 2014; 694 question/response pairs) On this basis, a typology of responses to questions is proposed – see Figure 1.

Figure 1:  
Typology  
of responses  
to questions.  
Source: Ginzburg  
*et al.* 2022, p. 86



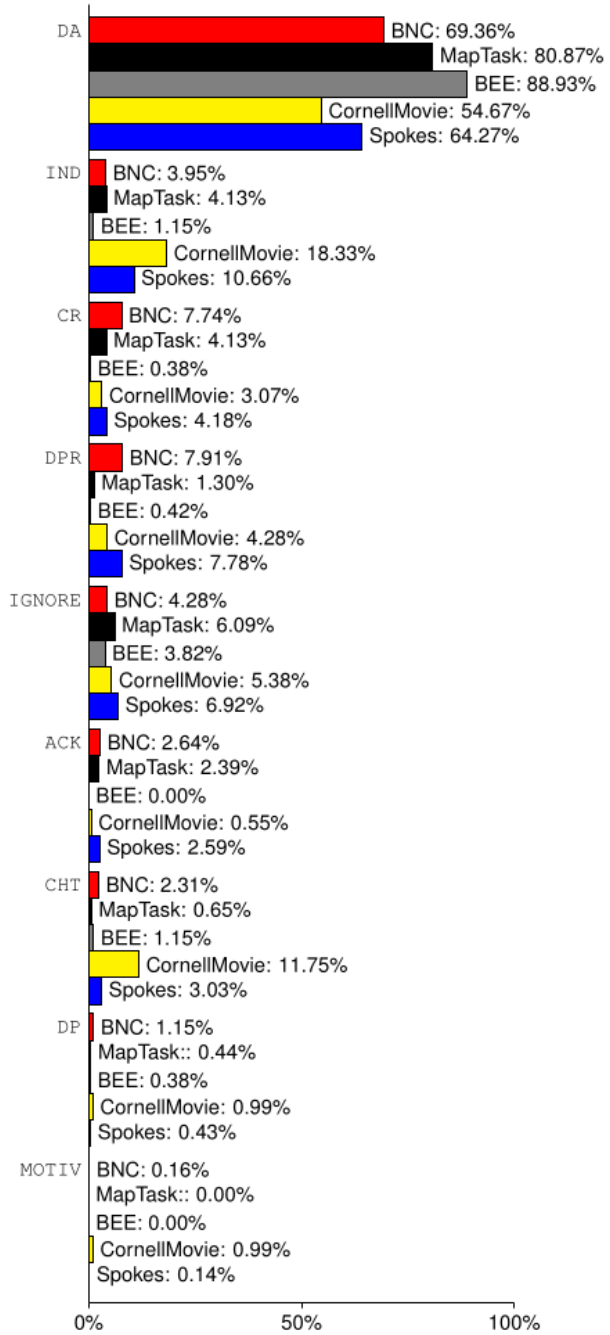
The two main categories of this typology are (1) *question-specific responses* (covering direct answers, dependent questions and indirect answers) and (2) *non-question-specific responses*. Direct answers (DA) provide an answer straightforwardly. For indirect answers (IND), one needs to infer an answer from the utterance. Dependent questions (DP) constitute a case where a question is provided as a response. What is more, the answer to the initial question (q1) depends on the answer to the query-response (q2). As for the non-question specific responses, we have: Clarification responses (CR) which address something that was not completely understood in the initial question (q1) and acknowledgements (ACK) wherein a speaker acknowledges that s(he) has heard and understood the question, e.g. *mhm*, *aha*, etc. Moving on to evasive question-responses, first we mention change-the-topic

(CHT). Instead of answering q1, the agent directly provides q2 and attempts to turn the table on the original querier. The original querier is pressured to answer q2 and put q1 aside. An IGNORE type of query-response appears when q2 relates to the situation described by q1 but not directly to the initial question. MOTIV is the type which addresses the motivation underlying asking q1. Whether an answer to q1 will be provided depends on a satisfactory answer to q2. DPR involves cases where the speaker states that it is difficult to provide an answer, points at a different information source, etc. or the speaker states that s(he) does not know the answer.

The corpus study revealed that for English the most frequent response classes in all four corpora are direct answers; the second most frequent class in the BNC is Difficult to Provide an Answer (DPR=7.91%), while in CornellMovie, the next biggest is indirect answers (IND=18.33%), whereas for the MapTask and BEE these are IGNORE (6.09% and 3.82% respectively). For Polish, the two most frequent classes of responses for Spokes are answers: direct ones (DA=64.27%) and – much smaller – indirect ones (IND=10.66%). The next two most frequent classes are DPR (stating that a person does not know the answer to the question, or it is difficult to provide one, DPR=7.78%) and utterances ignoring the question asked (questions and declaratives, IGNORE=6.92%). As illustrated in Figure 2 other classes are really rare – for MOTIV under 1% of the sample. This means that for certain response classes we have gathered very small numbers of examples. Such a result poses at least two challenges (as pointed out in the summary of Ginzburg *et al.* 2022). Firstly, how to collect more linguistic data for cross-linguistic testing? In the reviewed work, large English corpora were used but still certain classes of responses had small numbers of examples. The situation is even more challenging for languages lacking large or even hardly any speech corpora. Secondly, such a situation raises a serious difficulty when one thinks about potential applications of the corpus study with respect to dialogue interfaces. For such an application, machine learning should be used to acquire the response classification scheme (see Yusupujiang and Ginzburg 2022). This means that additional training and testing data are needed.

This brings us to a twofold motivation for designing QRGS. Firstly, to supplement the data from language corpora and open the way to

Figure 2:  
 Response types frequency  
 (BNC, n = 607;  
 BEE, n = 262;  
 MapTask, n = 460;  
 CornellMovie, n = 911;  
 Spokes, n = 694).  
 Source: Ginzburg et al.  
 2022, p. 93



apply machine learning approaches. Secondly, as not all languages have sizable linguistic corpora (see the disproportionate numbers for English and Polish in the aforementioned study) QRGS aims at closing this gap. This would pave the way for the cross-linguistic testing of the findings about the response space to questions (but not only).

The paper is structured as follows. Section 2 covers the main idea of QRGS and points at earlier work which it drew its inspiration. We also compare QRGS to selected, already existing crowdsourced solutions. Sections 3 to 6 present a series of QRGS evaluation studies. Starting from the pilot study where the effectiveness of the approach and correctness of the gathered data were checked, through questions concerning the non-native speakers' participation in QRGS, the role of game-like elements and the QRGS story theme. In Section 7, we describe a design of the crowdsourced evaluation module for QRGS. We end with the description of the part of QRGS data published as a part of the Erotetic Reasoning Corpus (Łupkowski *et al.* 2017). The summary gathers all the findings and points out aspects of QRGS that need further studies and improvements.

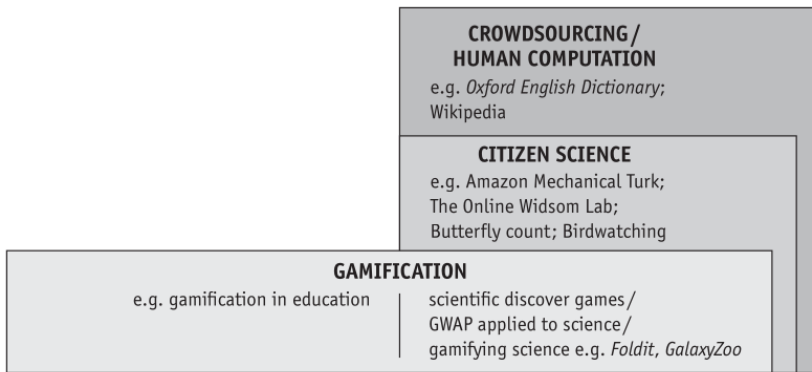
## QUESTION RESPONSES GENERATION SYSTEM – THE IDEA

2

The idea behind QRGS is to use *crowdsourcing* for relatively easy and effective collection of specific linguistic data. As such it may be identified as an example of a *scientific discovery game* (Cooper *et al.* 2010). A game of this kind is intended to help in processing large amounts of data obtained in scientific research. Two main tasks performed by human players in this case are mainly intelligent data analysis and classification tasks.

Scientific discovery games lie at the intersection of crowdsourcing, human computation and gamification – see Figure 3. Thus, we find methods and solutions known from these fields applied to solve given scientific problems. Typically, non-experts are employed to solve a given problem. As users perform the task in question in their free time and (usually) without gratification, it should be framed as relatively

Figure 3:  
A conceptual  
map of scientific  
discovery games.  
Source:  
Łupkowski and  
Dziedzic 2016,  
p. 129



simple and not time-consuming. Using game elements in a design is aimed at providing additional fun to the task, and also to motivate a user (e.g. with the points, achievements or leader boards).

A notable example of such a solution is Galaxy Zoo (Lintott *et al.* 2008). Galaxy Zoo was designed as a result of the huge amounts of astronomical data obtained from the Sloan Digital Sky Survey (SDSS). The problem for astronomers was to provide visual morphological classifications for nearly one million galaxies extracted from SDSS. Such a task is extremely difficult for current algorithms, and the work performed by small groups of experts had low efficiency (cf. Lintott *et al.* 2008). The idea of Galaxy Zoo is to provide users with a simple and brief tutorial and then allow them to perform classifications, using a very intuitive (symbolic) interface. Galaxy Zoo users are provided with photos of galaxies' from SDSS (the players are additionally motivated by the fact that most of the pictures have not been seen by anybody before them). Galaxy Zoo was so successful that it served as a template for analogous solutions for classification problems from other fields which are now hosted on the Zooniverse<sup>1</sup>.

Another interesting project of this kind is Foldit (Dsilva *et al.* 2019). Foldit is a perfect example of how a very difficult problem (3D modelling of protein structures) may be presented in the form of an easy to understand task – simple puzzle game.

From the field of linguistics it is worth mentioning such inspiring projects as PhraseDetectives (Chamberlain *et al.* 2008), which collects

<sup>1</sup><https://www.zooniverse.org/projects>.

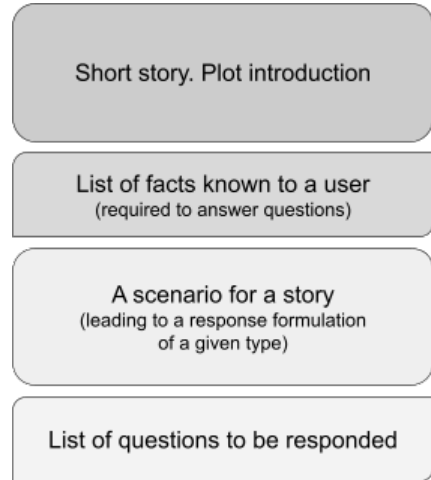


collaborative anaphoric decisions from online volunteers; Wordrobe (Venhuizen *et al.* 2013) which is a set of simple games developed to enable semantic annotation of the natural language data from the Groningen Meaning Bank (GMB); or RoboCorp (Dziedzic 2016) – the mobile game developed with the aim of annotation of the named entities retrieved from the Polish National Corpus (Przepiórkowski *et al.* 2011).

A direct inspiration for QRGS comes from the previous gamified solution related to questions and answers studies, which is the QuestGen described in Łupkowski and Wietrzycka 2015 and Ignaszak and Łupkowski 2017. QuestGen is a game-like system, in which players generate questions of a specific form while solving a detective game. In the game, two randomly chosen players are engaged in solving a detective puzzle. One of them plays as the Detective, while the other is called the Informer. The aim for the Detective is to solve the presented puzzle by questioning the Informer. Each story in the game has two formulations (one for the Detective and one for the Informer), containing all the additional data necessary to solve the puzzle. Each story should be solved within a given time limit. For each story the players switched roles, from the Detective to the Informer and *vice versa*. Players were not supervised in any way, they were just playing the game. Crucially, stories' plots were formulated according to erotetic search scenarios, a tool developed within Inferential Erotetic Logic (Wiśniewski 2013). Thanks to this, each story has only one correct solution and a normative way to reach it (pointed out by the underlying erotetic search scenario). Overall, 116 game transcripts from 40 players were collected. The general solution statistics for the study sample (all six stories) is the following: 91 solutions are correct, out of which 44 are normative, i.e. solved exactly according to the e-scenario underpinning a given story. In 18 cases, Detectives provided incorrect solutions and in 7 they did not provide any solution (mostly due to time constraints) – see detailed discussion in Ignaszak and Łupkowski 2017.

QRGS relies on a very similar schema. A QRGS user is asked to read the simple story and impersonate its main character. As the story unfolds, a user is confronted with four questions and (s)he is expected to answer them in the way the main character would do that. As the story unfolds, a user is confronted with questions related to the story

Figure 4:  
A general QRGS schema



and (s)he is expected to answer them in a way the main character would in the given context – see Figure 4 for a general QRGS schema. Stories which are prepared for QRGS to this point are presented in detail in Section 3 and Appendices A, B and C.

Here we should also mention yet another crowdsourced solution for gathering question-response pairs. The motivation for the solution also comes from the corpus study presented earlier and it is aimed at addressing the challenge of characterising the response space to questions in a low-resource language – Uyghur. The early design is presented in Yusupujiang and Ginzburg 2020 and Yusupujiang and Ginzburg 2021. Initial studies and results are discussed in Yusupujiang and Ginzburg 2022. The paper presents a Uyghur dialogue corpus based on a chatroom environment (using the Rocket.Chat implementation). The Uyghur Chat-based Dialogue Corpus (UgChDial) is divided into two parts: (1) Two-party dialogues and (2) Multi-party dialogues. It consists of 25 chat sessions, with 6 participants, resulting in 1,581 question-response pairs. The sessions were based on different scenarios and topics. The analogue to QRGS are role-playing scenarios, which require participants to act according to certain situations (such as police vs. criminal, debtor vs. debtee, sales person vs. a customer with complaint). This is aimed at retrieving evasive or cooperative responses from users.

## PILOT STUDY. PROOF OF CONCEPT

3

In this section, we present the pilot study of QRGS. The study was aimed at answering the following research questions.

1. How *effective* is QRGS in terms of data gathering – how many question/response pairs (Q-Rs) will be gathered and how long will it take?
2. How (linguistically) *interesting* are the gathered responses? Namely, will the responses generated to presented questions differ between subjects? Will they be comparable to responses that may be observed during a dialogue?
3. Are the gathered responses *correct*? I.e. are they of the type which is expected for a given scenario for the story?

*Tools and materials*

3.1

For the study, two stories were prepared: “The Bomb” and “The Party”. We describe them in detail below. For each story, we firstly present a user with the introductory plot including the facts known to the user. After that, four scenarios are presented to a user along with the questions (the same four questions are used for all scenarios). The task of a user is to immerse into the story and provide responses to the presented questions, which will be formulated in a manner appropriate to a given story and the current scenario).

The first story is entitled “The Bomb” and was adapted from the previous studies related to questions and question answering (Urbański *et al.* 2016a). The plot presented to a participant is the following.

A bomb was planted in the main train station of Nibyjunkcja. You are the chief of security at the train station where the bomb was planted. After checking the security cameras you have established the following facts:

1. The bomb was planted under the third pillar.
2. The bomb has the size of a shoe-box.

3. It was planted by a tall guy dressed in a red T-shirt.
4. It was planted between 8:00 and 8:30 A.M.

The first scenario for the story is such that a subject should provide a direct answer to the questions asked. It is entitled “The coordinator of the sapper unit”.

You are approached by the coordinator of the sapper unit who is trying to establish which wire to cut in order to disarm the bomb. You are *obliged to be truthful and give direct and precise answers* to his questions. Please answer the following:

1. Do you know where the bomb was planted?
2. How big is the bomb?
3. Can you describe the suspect?
4. Can you tell me when the bomb was planted?

As a result we should obtain four direct answers (DA) to the introduced questions.

The second scenario for the story is: “A trusted journalist”. For this we are expecting indirect answers (IA). To encourage a participant to provide such responses the following lead is used.

You are approached by Frank, a journalist for the local “Nibyjuncja Today”. You have known Frank for a long time and trust him. He wants to gather some news about the situation on the station. Given that the investigation is in progress you cannot give Frank direct information. Nonetheless, since you trust him, *try to provide truthful information but in an indirect manner*. Please answer the following questions of Frank. / Here the same set of questions is used as for the first scenario, p. 222. /

The next scenario is “A journalist you do not trust”, which is aimed at retrieving evasive and polite responses.

You are approached by a journalist you do not know. His ID indicates that he came from the capital and works for the big journal “NBJ News”. You do not trust him. However, you are obliged to answer his questions in order to avoid problems with the press. Please answer the journalist’s questions

*in such a way that he will understand that you do not want to answer his questions (be polite...).* Please answer the following. / Here the same set of questions is used as for the first scenario, p. 222. /

And the last one is entitled “A random guy” (for evasive and impolite responses).

You are approached by a random guy from the crowd surrounding the scene. He tries to ask you some questions. Please answer them in such a way that he will understand immediately that you *do not want to answer* his questions (you do not have to be extremely polite, however you should not lie, or simply answer using “no comments”). Please answer the following. / Here the same set of questions is used as for the first scenario, p. 222. /

“The Party”. The second story considers inviting people to a party. It also has four scenarios. The plot is introduced in the following paragraph.

Imagine that you are organising a party next Saturday. You want to invite just several close friends: Ann, John, Frank, Alice and Bill. The party is on Saturday and starts at 8 P.M. You would like it to end around midnight. You plan a barbecue and beer in the garden.

As in the previous case, four scenarios, each aimed at a different category of responses obtained were designed. “Alice” for direct answers (DA).

In a shop, you are approached by Alice. She is already invited to the party and has accepted the invitation, so you can *openly and directly* answer her questions. Please answer the following questions asked by Alice:

1. How many people will there be at the party?
2. Is Ann invited?
3. Will there be any alcohol at the party?
4. When do you want to start?

“Helen” for evasive answers (polite).

In a shop, you are approached by Helen. She is your neighbour and somehow got to know about the party. You do not want to discuss any details with her so answer her questions in such a way that she will know that *you do not want to answer them* (still do be polite, she is your neighbour after all). Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario. /

“Willy” for evasive answers (impolite).

While coming back from work you are approached by little Willy, your neighbours’ son. He tries to ask you some questions. Please answer them in such a way that he will understand immediately that you *do not want to answer* his questions (you do not have to be extremely polite, however you should not lie). Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario, p. 223. /

And the last scenario is “John” for indirect answers (IA).

During the evening John calls you. You are in one room with your friend, who does not know about the party. John is asking some questions. Please answer them in an *indirect* manner so that your friend will not get any idea concerning the party. Please answer the following. / Here the same set of questions is used as for the first “The Party” scenario, p. 223. /

The summary of expected question responses types to the different formulations of stories is presented in Table 1.

### 3.2

#### *The Procedure and Participants*

Stories and questions were presented online with the use of the Google Forms platform (each scenario for a story separately). Only text was presented, no additional images were included to supplement stories. Instructions for the participants were the following:

Story	Scenario	Expected response
“The Bomb”	The coordinator	DA
	A journalist (trusted)	IA
	A journalist (untrusted)	Evasive (polite)
	A random guy	Evasive (impolite)
“The Party”	Alice	DA
	Helen	Evasive (polite)
	Willy	Evasive (impolite)
	John	IA

Table 1:  
Expected question responses to the different formulations of stories

Below you will find a short story and 4 questions for it. Please try as best as you can to get into the character and write how you would answer the questions below in real life. The speed of completing the task will not be measured, so please take as much time as you need.

Invitations for participants (each participant for each variant of a story) were sent out *via* social media. No information was collected about the participants (which is a common practice for crowdsourcing tools), however the invitations were intentionally sent to people without experience in linguistics and with a high level of English language proficiency. 25 participants took part in the study. The data collection lasted from the 1st to the 5th May 2018.

### Results and data validation

3.3

**Effectiveness.** Overall we gathered a sample of 100 Q-R pairs generated by 25 participants in just five days. The summary of generated responses is presented in Table 2. One may conclude that QRGS is effective when it comes to the numbers of gathered responses and the data collection time. This is mainly due to the fact that the task for a participant is not very demanding and the data collection itself does not require any supervision from the researcher.

Let us now take a closer look at the variety of the gathered data. In order to be useful for the intended use, the question responses retrieved for one question should have different formulations. The QRGS

Table 2:  
Number  
of responses  
generated  
for each  
QRGS story

Story	Participants	Responses generated	Response type
Bomb 1	4	16	DA
Bomb 2	4	16	IA
Bomb 3	5	20	EAP
Bomb 4	4	16	EAI
Party 1	2	8	DA
Party 2	2	8	IA
Party 3	2	8	EAP
Party 4	2	8	EAI
Sum	25	100	–

data would not be interesting if we would obtain e.g., 50 “Yes” responses to the question “Do you know where the bomb was planted?”. Fortunately this is not the case. We observe a wide variety of the retrieved question responses. Consider the following examples.

For the “Bomb” history and scenario “Untrusted journalist” and question *Do you know where the bomb was planted?* we have responses such as the following (in all examples we preserve the original spelling):

- This information is available for me.
- Where would you plant such a bomb?
- All stations are being monitored. We have the data from the cameras – therefore we will be able to localise any unusual behaviour.
- Yes.
- There are some clues to figure out where the bomb is. It is probably somewhere nearby.

And for the same story, but the scenario “Trusted journalist” and question *How big is the bomb?*:

- The bomb could have been carried by a single person in a handbag
- Did you finally manage to reduce the size of space occupied by your precious collection by throwing out the unnecessary stuff?
- Let’s say that you can carry it in a shopping bag.

One may observe that the responses generated by our participants vary with the respect to complexity, length and style. Thus we are



Response category	Generated	Correct	(% corr)
DA	24	24	100%
IA	24	13	54%
EAP	28	18	64%
EAI	24	22	92%
All	100	77	77%

Table 3:  
Summary of responses' correctness  
with respect to categories

gathering responses which are close to the natural language dialogue outcomes. This also suggests that, to a large extent, our participants were able to immerse into the storyline presented and answer questions suitable to the plot.

**Correctness.** Naturally the most important question is whether these generated responses were of an expected type – i.e. were they correct? This aspect is very important as the data gathering with QRGS is not supervised. A high percentage of correct answers is needed for the data to be useful for future applications. To answer this question, the responses were manually evaluated by two researchers. The aim was to check whether the actual responses given by participants fit into the expected categories.

Each response was tagged independently by two annotators using the following tagset: DA (direct answers), IA (indirect answers), EAP (evasive polite), EAI (evasive impolite), OTHER (for cases not matching the listed categories). Inter-annotator agreement was then calculated with the use of Kappa statistics. In what followed, a final tag was assigned to each response as a result of discussion between annotators. This tag was then compared with the intended category for a given response.

For the reported study the agreement between both raters was measured using Cohen's kappa coefficient. This was established using the R programming language (version 3.5.0) and the irr library. Cohen's kappa was 0.775 (which indicates the substantial agreement between raters, see Viera and Garrett 2005).

The manual evaluation shows that 77% of responses are in line with the predictions – see details in Table 3.

**Error analysis.** Participants of this study had no problems with providing direct answers. All of the gathered DA responses were

correct. As for indirect answers (IA) we observe a common mistake, which is providing a DA instead of an IA, like in the following example:

A: Do you know where the bomb was planted?

B: *Yes, somewhere in the station.*

This constitutes 10 of the 11 observed errors. Only one error was that instead of an IA an evasive answer was provided.

A: How many people will there be at the party?

B: *I really enjoy spending time with my close friends.*

Let us now take a closer look at evasive responses. All the mistakes in this case were that instead of an evasive answer a direct one was provided (however, these were partial answers). This is exemplified in the following:

A: Can you tell me when the bomb was planted?

B: *Certainly today.*

Interestingly, more errors for evasive responses were observed for the polite condition than for the impolite condition – this suggests that for the participants the impolite condition was easier to formulate such responses.

**Summary.** QRGS proved to be a simple and effective crowdsourcing tool for gathering interesting data. The task is not demanding for a user and is thus very quick to complete. QRGS needs no supervision on the level of data collection. Also, the data correctness in our study is satisfactory. It is worth stressing that incorrect responses (i.e. the ones that do not match the expected type for a given scenario) are not useless for future applications. Manual re-annotation leads to their classification to the appropriate type.

The pilot study results presented in this section led to further research questions and potential improvements for QRGS. Firstly, we have not gathered any information concerning our participants. For a QRGS evaluation it would be useful to learn whether language proficiency matters for data generated with QRGS; in particular, whether we would observe differences between native and non-native speakers. Another question addresses the level of game-like elements involved in QRGS – would it be better to supplement QRGS stories with

graphics? Last but not least, it is an open question whether the type of story plot for QRGS matters for the results. We address these questions in the following sections.

## NATIVE VS NON-NATIVE ENGLISH SPEAKERS 4

In this section, we present a study focused on the research question whether we would observe any differences in QRGS outputs for native and non-native English speakers groups. Given that in the pilot study we did not gather any data concerning participants, this question remains open. The answer is important for potential QRGS applications.

### *Materials* 4.1

For the purpose of this study, two previously written QRGS stories were used (“Bomb” and “Party”). Also, two new ones were prepared. These were “The Epilepsy” and “The Secret Santa”. The first one is a story that your co-worker, Anna, has just had an epilepsy attack and you helped her and called an ambulance. The second story revolves around you and your friends having decided to organise a Secret Santa event this year and you considering different ideas for presents. New stories were prepared exactly in line with the first two. After a short introduction of the situation and of the known facts, a user is presented with four scenarios along with questions – each scenario formulated in such a way that it leads to different types of responses (direct ones, indirect ones, evasive polite and evasive impolite). The complete stories along with their corresponding scenarios are presented in Appendix A.

### *Procedure* 4.2

The study was conducted *via* the Internet using the Google Forms platform. Participants were presented with one short story each and asked to answer 4 questions. Participants were asked to “enter into” the situation and empathise with the assigned role and provide written answers as if they were responding directly to the character from the

story. It was made clear in the instructions that no time limit for the task completion was assumed. After answering all the four questions, the participant was asked to answer several demographic questions (covering age, gender, education, native language and for non-native English speakers their English proficiency level).

### 4.3

#### *Study group*

The group consisted of 49 participants, of which 28 were female, 19 were male and 2 preferred not to reveal their gender information. Participants were recruited *via* social media. The average age was 32.02 (SD=11.67, min=15, max=57). The declared education level was the following: doctoral degree: 7; university degree: 27; high school diploma or equivalent degree: 9; less than high school diploma: 6. Most importantly 31 participants were native speakers of English and 18 were non-native speakers (12 Polish; 2 Czech; 1 Spanish; 1 Swedish; 1 Azerbaijani; 1 Arabic). The declared English proficiency level for the non-native speaker group was the following: A2 (Elementary): 1; B1 (Intermediate): 2; B2 (Upper Intermediate): 4; C1 (Advanced): 7 and C2 (Proficient): 4.

### 4.4

#### *Results and data validation*

**Effectiveness.** The data were collected during December 2018 and March–May 2019. Overall participants generated 196 responses. 124 in the native speakers group and 72 in the non-native group – see details in Table 4.

Table 4:  
Summary  
of responses'  
correctness  
with respect  
to groups  
and categories

Response type	Native	(% corr)	Non-native	(% corr )
DA	48	96%	20	100%
IA	12	42%	12	58%
EAP	36	36%	28	28%
EAI	28	50%	12	12%
All	124	63%	72	62%

**Variety.** Firstly, we observe that the generated responses are interesting and differ between participants. Examples are provided below. For the “Bomb” story (scenario the unit coordinator) and question: *Do you know where the bomb was planted?* we have for example:

- Under the third pillar in the Nibyjunkcja main train station.
- In the main train station of Nibyjunkcja, at the base of the third pillar.
- In the main train station of Nibyjunkcja.

For the “Secret Santa” story (DA scenario) and question: *What are we giving to Joe?*

- Craft Beer Brewing Kit.
- We’re giving him the Craft Beer Brewing Kit.
- Craft Beer Brewing Kit.
- the craft beer brewing kit.

We also find responses which are carefully prepared and much longer than the presented ones.

*What are we giving to Joe?:* I don’t know much about home brewing, so this would be a bit difficult. I’d try to reference something that ‘hops’. My best idea so far is to talk about a trip to the zoo. My family went to the zoo last week. We watched the kangaroos for ages and my daughter insisted on hops, hops, hops to get around after that. We found a nice kit at the gift store that will allow us to make our own hoppy creature. It’ll be especially good for taking with us to enjoy at BBQs.

*How much is the contribution rate?:* If I were in the US, I’d say something about Hamilton, the musical. (Alexander Hamilton is on the 10 Dollar bill.) If I were in Australia, I would say something about the Wattle tree or quote The Man from Iron Bark (Both the Wattle and Banjo Patterson are on the A 10 Dollar note).

As in the case of the pilot study we may conclude that QRGS data are interesting and reminiscent of responses provided in spontaneous conversation.

Table 5:  
Length of responses  
(number of characters)  
generated in QRGS

Group	Mean	SD	Median	Min	Max
Native	38.34	57.75	26.50	1	420
Non-native	30.56	29.79	19.50	2	152

**Correctness.** For the data evaluation a procedure analogous to the one described in Section 3.3 was applied.

Two annotators were engaged in the evaluation. The agreement between both raters was measured using Cohen's kappa coefficient (established using the R programming language (version 3.5.0) and the irr library). Kappa for the native speakers group = 0.717, and for the non-native speakers group = 0.639. Both results indicate substantial agreement between raters (see Viera and Garrett 2005).

The general correctness of the respondents in the group of native speakers was 63% and for the non-native speakers group it was 62%. One may conclude that for the correctness factor of the gathered data these groups do not differ. The detailed summary is presented in Table 4. As expected, providing a Direct Answer to a question was the easiest task, with almost 100% accuracy in both groups.

We also decided to take a closer look at the length of the generated responses in order to check whether they differ between groups. The intuition behind this step is that the length of a response (in the numbers of characters used) provides a rough (quantitative) indication of how elaborate the response is. One may expect that the native group would provide longer, more elaborate responses.

The length of responses for the groups is presented in Table 5 and Figure 5.

The Wilcoxon Test shows that there are no statistically significant differences between the groups ( $W = 4504.5$ ,  $p = 0.9168$ ). The responses provided by QRGS users do not differ between groups of native and non-native English speakers. Their correctness is at a similar level. Also, the average number of characters per response indicates that responses were similarly complex when it comes to formulation. Such a result is promising for future QRGS implementation for popular languages (such as English). QRGS may be used to gather data for such languages even if the access to the group of native speakers is limited. Naturally, this may not be easily generalised for other languages and needs further testing.

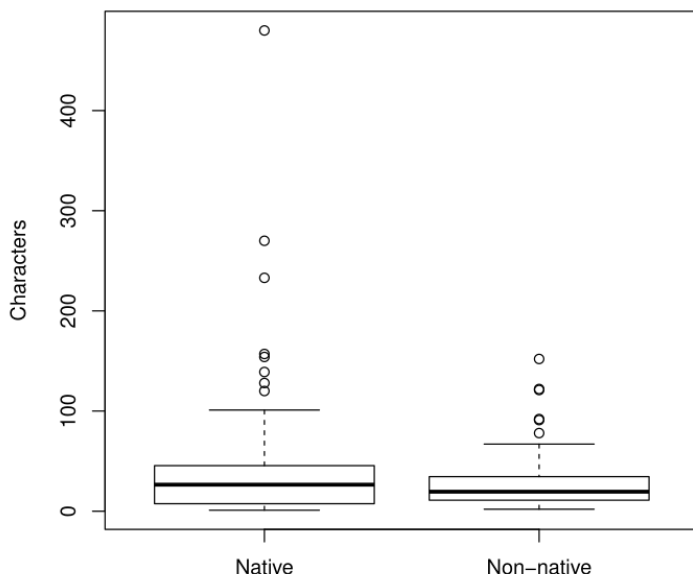


Figure 5: Comparison of number of characters used in responses generated by group of native and non-native English speakers

## GRAPHICAL VS TEXTUAL VERSION

5

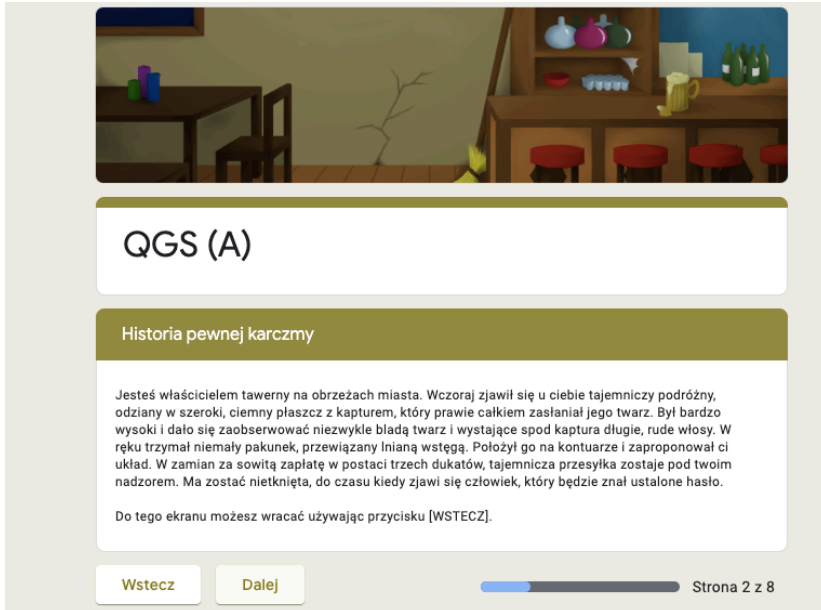
The following section describes a study where we asked the research question whether QRGS should involve more game-like elements, especially graphical ones. The intuition would be that a more game-like system will stronger immerse participants. The more immersed the participant, the better (i.e. more natural and correct) responses provided for QRGS stories. Thus, we designed a graphical version of QRGS for the experimental group, while the control group used the already tested textual one. We studied the differences in outcomes in terms of correctness of the data, response length and the self-declared engagement of users.

### *Materials*

5.1

For the purpose of this study, a new QRGS story was prepared in Polish. It is entitled “The Tavern” and tells a story of a tavern owner who is asked for a favour – storing a mystery object for an unknown person. The complete story along with its corresponding scenarios is presented in Appendix B.

Figure 6:  
Textual version  
of QRGs. English  
translation  
of the story  
in Appendix B



As mentioned above, two versions of QRGs were prepared. First, the traditional one, i.e. textual (as presented in Figure 6). The header of a questionnaire was supplemented with one simple graphic presenting the inside of a tavern. The second version was a graphical one with the style inspired by RPG games. The story was presented step by step with the appropriate illustrations (Figure 7). Also, the characters from the story were presented in the visual form (Figure 8). It is worth stressing that the text presented in both versions was identical.

In order to assess the engagement level of the participants, we employed the shortened version of the IMUW questionnaire. IMUW (Wasielewska and Łupkowski 2022) is a questionnaire based on the Polish adaptation (Strojny and Strojny 2014) of the immersion questionnaire (Jennett *et al.* 2008). It measures self declared engagement into task performance. The full IMUW consists of 25 items. For the purpose of the study we, prepared a 10 item short version (as the IMUW reliability study reports that it is a one factor questionnaire). Below, we present this IMUW version with the English translation of items.

1. W jakim stopniu zadanie podtrzymało Twoją uwagę? / *To what extent did the task hold your attention?*



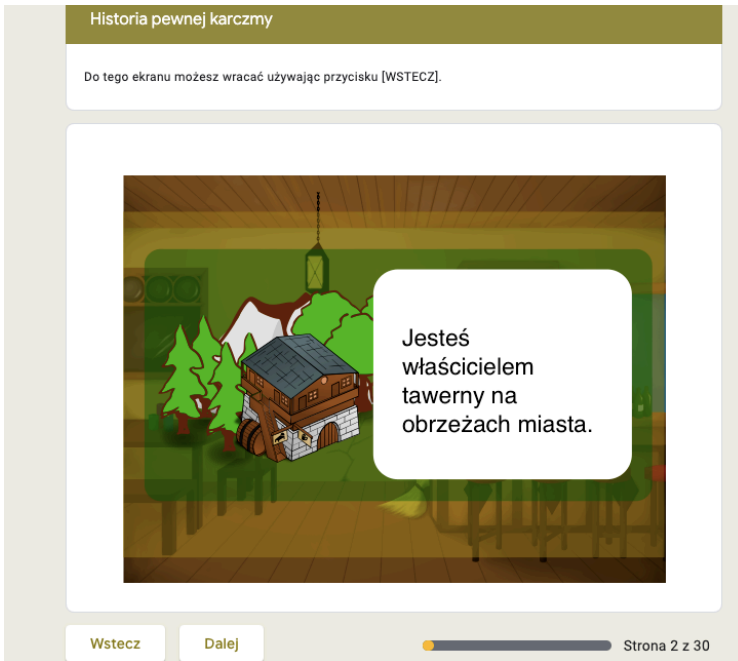


Figure 7: Graphical QRGS version. The story unfolds step by step and is illustrated. The panel says “You are the owner of a tavern in the suburbs”. Full story in English in Appendix B

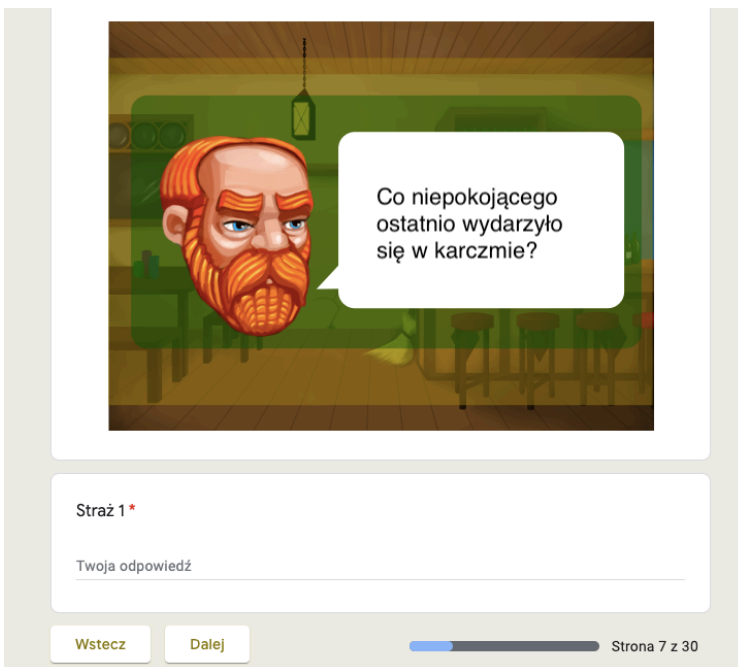


Figure 8: Graphical QRGS version. Characters from the story are presented, and dialogues are simulated. The panel says “What was the worrying thing that happened at the tavern?”. Full story in English in Appendix B

2. W jakim stopniu odczuwałeś(aś), że jesteś skupiony(a) na zadaniu? / *To what extent did you feel you were focused on the task?*
3. Jak dużo wysiłku włożyłeś(aś) w wykonanie zadania? / *How much effort did you put into playing the game?*
4. Czy odczuwałeś(aś) w którejkolwiek chwili potrzebę przerwania wykonywania zadania i zobaczenia, co się dzieje wokół? / *Did you feel the urge at any point to stop performing the task and see what was happening around you?*
5. W jakim stopniu odczuwałeś(aś), że zadanie jest czymś, czego raczej doświadczasz niż po prostu czymś, co robisz? / *To what extent did you feel that the task was something you were experiencing, rather than something you were just doing?*
6. W jakim stopniu czułeś(aś) się emocjonalnie zaangażowany(a) w zadanie? / *To what extent did you feel emotionally engaged in the task?*
7. W jakim stopniu byłeś(aś) zainteresowany(a) tym, jak potoczy się fabuła czytanego przez Ciebie tekstu? / *To what extent were you interested in seeing how the presented story plot would progress?*
8. W jakim stopniu podobał Ci się poziom artystyczny tekstu? / *To what extent did you enjoy the presented text?*
9. Jak dużą czerpałeś(aś) przyjemność z wykonywania zadania? / *How much would you say you enjoyed performing the task?*
10. Czy chciałbyś(aś) wykonać zadanie jeszcze raz? / *Would you like to perform the task again?*

## 5.2

### *Procedure*

The study was conducted online with the use of Google Forms. Participants were invited to take part in the study *via* a link on the social media pages. The link led to the page where a participant was randomly assigned to one of the groups. Participants received necessary information about the study and provided their agreement to take part. After that, they were presented with the story followed by four scenarios with questions. Next, they filled out the IMUW questionnaire and provided basic demographic data.

*Study group*

5.3

70 participants took part in the study. 35 in group A (textual QRGS version), aged 18-31 (mean 22.43; SD = 3.19), 62.9% women. 35 in group B (graphical QRGS version), aged 19-41 (mean 24.85; SD=5.67), 54,3% women. All participants were native Polish speakers.

*Results and data validation*

5.4

The data was collected from the 10th of March 2019 till the 22nd of March 2019. Overall 1,120 responses were collected. The variety of responses was satisfactory, as observed for previous studies in English.

**Correctness.** To assess response correctness, we randomly chose 100 Q-R pairs from group A and 100 from group B. A procedure analogous to the one described in Section 3.3 was applied. For this study, each response was tagged independently by three annotators, thus inter-annotator agreement was controlled for with the use of the Fleiss kappa coefficient (established using the R programming language, version 3.5.0, with the irr package). Fleiss’ kappa was 0.504 for group A and 0.575 for group B. As for percentage of correct answers, we got 49% for group A and 59% for group B – details are presented in Tables 6 and 7. We observed a small advantage in the case of the graphical QRGS when it comes to providing responses according to the expected type.

**Length.** As in the case of native/non-native speaker study we decided to check the length of responses provided in both groups. The length of the responses for the groups is presented in Table 8 and Figure 9.

Response type	Generated (A) <sup>a)</sup>	Correct (A)	(% corr)
DA	25	20	80%
IA	25	13	52%
EAP	25	7	28%
EAI	25	9	36%
All	100	49	49%

Table 6:  
Summary of responses’ correctness with respect to categories for group A (textual)

<sup>a)</sup> Subset of the whole sample.

Table 7:  
Summary of responses' correctness with respect to categories for group B (graphical)

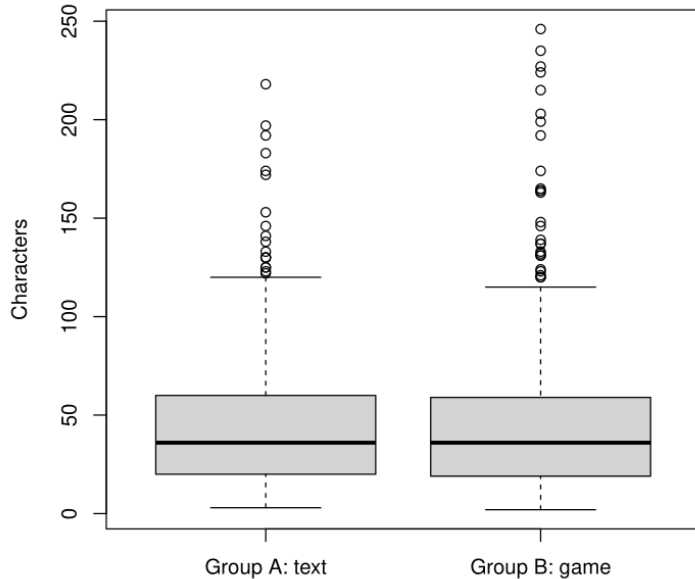
Response type	Generated (B) <sup>b)</sup>	Correct (B)	(% corr)
DA	25	23	92%
IA	25	8	32%
EAP	25	10	40%
EAI	25	18	72%
All	100	59	59%

<sup>b)</sup> Subset of the whole sample.

Table 8:  
Length of responses (number of characters) generated in QRGS

Group	Mean	SD	Median	Min	Max
A	45.10	33.91	36.00	3	218
B	44.98	37.53	36.00	2	246

Figure 9:  
Comparison of the number of characters used in responses generated by groups A (textual QRGS) and B (graphical QRGS)



The Wilcoxon Test shows that there are no statistically significant differences between the groups when it comes to the length of the responses ( $W = 160198$ ,  $p = 0.5302$ ).

**Engagement.** The Cronbach alpha of IMUW for this study was 0.86 for group A and 0.83 for group B. Hence, we can confirm that

Group	Mean	SD	Median	Min	Max
A	37.17	7.39	38	20	50
B	31.80	6.68	32	20	45

Table 9: IMUW results (declared engagement in the task) for groups A (textual QRGS) and B (graphical QRGS)

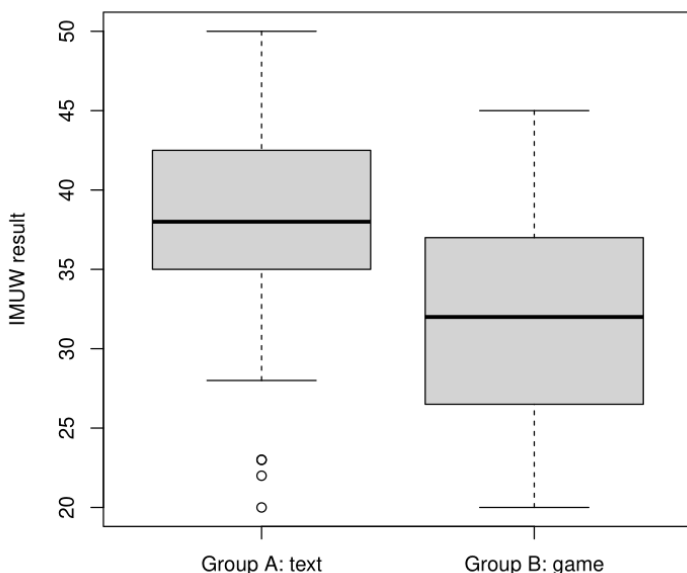


Figure 10: Comparison of IMUW results (declared engagement in the task) for groups A (textual QRGS) and B (graphical QRGS)

the reliability of the tool used was high. IMUW results are presented in Table 9 and Figure 10.

The Wilcoxon Test shows that the difference between group A and group B is statistically significant ( $W = 874.5$ ,  $p = 0.002106$ ). We may conclude that the textual version of QRGS was more engaging for our participants than the game-like, graphical one.

This study indicates that a step towards more game-like solutions for QRGS is not necessary. In terms of correctness of the gathered data and response length, we do not observe any apparent differences between groups. What is interesting is the result related to the self reported engagement into the task. Smaller engagement for the graphical version of QRGS may suggest more distracting factors exist for this version. This may be also the result of the fact that in this version the story unfolds more slowly and the whole task takes more time.

Definite answers on these issues require further investigation. However, on the basis of the results obtained already we can say that the textual version of QRGS is still a good option to be used – especially due to the simplicity of the design and implementation.

## 6 PLOT FORMULATION

In the following study, our aim was to test QRGS in yet another respect. Namely, whether the style of the plot of the story used matters for correctness. The intuition behind this question is that certain types of stories may be more immersive or more appealing to users and thus result in more correct responses being generated. That is why we decided to design two new QRGS stories in Polish, one of which is a detective story in which a participant is lured into the crime-solving plot. The second one is more neutral as it concerns organisation of an engagement surprise party.

### 6.1 *Materials*

For the sake of the study two QRGS stories were prepared: “Jewellery theft” and “Engagement”. The first one tells the story of a bold theft of old jewellery from a nobleman’s home. A participant takes part in the interrogations to find the culprit. Thus, the questions to be responded to in QRGS concern the following: what did the thief look like? What did he use to carry the stolen goods? How did the thief manage to escape the home? What time did the theft take place? The second story concerns an engagement surprise party. A participant plays the role of a friend asked to book the restaurant. Questions to be responded to cover the time of the party, number of people to be invited or planned surprises. Stories and their scenarios are presented in Appendix C.

### 6.2 *Procedure*

The study was conducted online using Google Forms. Two separate forms were prepared for the two stories. Participants were invited to take part in the study *via* a link on the social media platforms. The link

led to the page where a participant was randomly assigned to one of the stories. Participants received the information about the study and provided their agreement to take part. After that, they were presented with the story followed by four scenarios with questions. At the end, they provided basic demographic data.

### *Study group* 6.3

Overall, 199 participants took part in the study. The “Engagement” story form (group A) was filled out by 101 participants, including 90 women and 11 men. The participants were between 17 and 45 years, and the mean was 21.84 years ( $SD = 5.13$ ). The version with the story of jewellery theft (group B) was filled out by 98 people, of which 88 of individuals were women and 10 are men. The age of the participants ranged from 17 to 45 years with an average of 22.48 years ( $SD = 5.53$ ).

### *Results and data validation* 6.4

Data was gathered from May 2019 till January 2020. We collected 3,184 responses: 1,616 responses to the first story and 1,568 to the second one. The variety of responses was satisfactory, as observed for previous QRGS studies.

**Correctness.** The correctness check covered the whole gathered sample. We used the same procedure as in previous studies. Fleiss’s kappa (for three annotators) was 0.502 for group A and 0.527 for group B. As for the percentage of correct answers, we got 54% for group A and 57% for group B – see the details in Tables 10 and 11. Thus, when it comes to correctness the plot formulations are very similar.

**Length.** As in the case of previous studies, we decided to check the length of responses provided in both groups. The length of responses for the groups is presented in Table 12 and Figure 11.

The Wilcoxon Test shows a statistically significant difference between groups ( $W = 1525410$ ,  $p < 0.001$ ). Responses gathered for the detective-like story were significantly longer than the ones for the story about the surprise party.

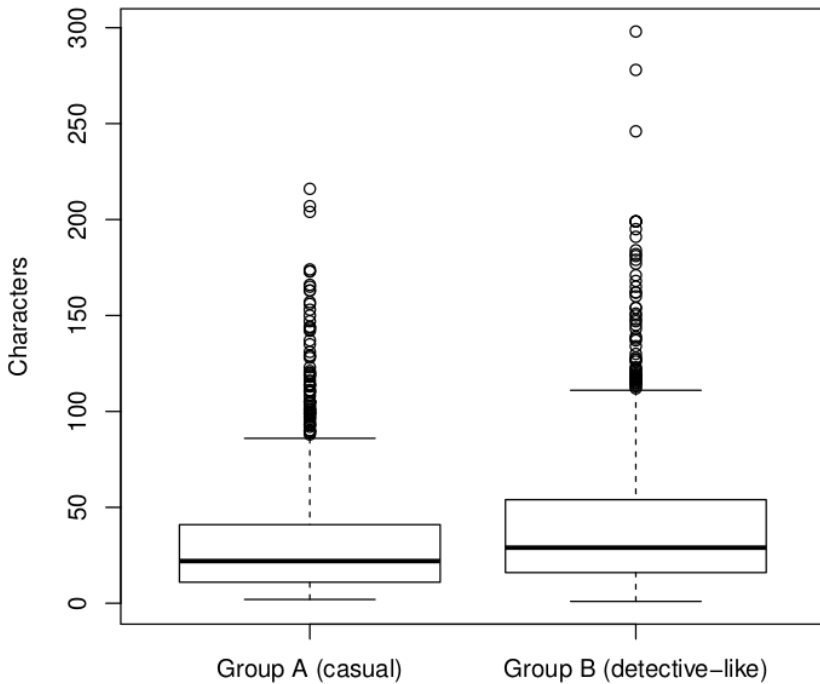
Table 10:  
Summary  
of responses'  
correctness  
with respect  
to categories  
for group A  
(*Kradzież  
biżuterii*)

Response type	Generated (A)	Correct (A)	(% corr)
DA	392	388	99%
IA	392	78	20%
EAP	392	225	57%
EAI	392	154	39%
All	1,568	845	54%

Table 11:  
Summary  
of responses'  
correctness  
with respect  
to categories  
for group B  
(*Zaręczyny*)

Response type	Generated (B)	Correct (B)	(% corr)
DA	404	410	99%
IA	404	144	36%
EAP	404	199	49%
EAI	404	186	46%
All	1,616	930	57%

Figure 11:  
Length  
of responses  
(number  
of characters)  
generated  
in QRGS





Group	Mean	SD	Median	Min	Max
A	30.09	28.65	22	2	216
B	39.65	33.75	29	1	298

Table 12:  
Length of responses  
(number of characters)  
generated in QRGS

The results indicate that no apparent differences are observed between two different story topics when it comes to the correctness factor. The difference is observed in terms of length. For the detective-like stories, provided responses were longer. In consequence we may conclude that detective-like stories are recommended for QRGS if one wishes to obtain longer responses. These also seem to be easier to plan and write. However, as the differences in correctness are very small, the choice of topics to be used for QRGS stories is open.

## QRGS EVALUATION MODULE

7

The studies described in previous sections revealed a potential weakness of the QRGS framework, namely the need for a manual data check (after they have been gathered). The process of data gathering needs no supervision. Data correctness is also satisfactory (especially for selected types of responses, like DA). However when one thinks about the potential use of the QRGS for supplementing carefully collected corpus data, it certainly requires additional control.

To deal with this issue, we decided to design, implement and check the evaluation module for QRGS which also uses a crowdsourcing mechanism. In this scenario, a user’s task is not only to generate new responses but also to evaluate selected responses previously provided by other players. As the generation phase is not demanding and is rather short, we believe that adding the evaluation phase to QRGS would not be troublesome for a user.

### *Two evaluation module designs*

7.1

For the evaluation module, we assume that a user has previously read the story and generated responses to provide questions. In the evaluation module, the user is asked to perform only a simple matching

task. The task has two versions: (EM-A) matching story characters to responses, and (EM-B) matching one of the response categories to a provided response. After completing the task, a user is asked to assess how certain s(he) is about the solution proposed (the higher the assessment, the more certain the user is). This is heavily inspired by the Wordrobe (Venhuizen *et al.* 2013) – in this system it was observed that it results in better user performance. For future QRGs applications EM-A or EM-B may be used separately or together (to add more task diversity).

**EM-A.** A user is presented with the instruction that s(he) should match four characters to the four responses given and for each choice declare how certain s(he) is about the match. Each character has a short description of the type of response it provides (according to the story plot) – see Figure 12.

**EM-B.** A user is presented with a question/response pair (in a form known from chat applications) below which the one-choice list of response types is presented. The user chooses one of the answers and declares how certain (s)he is about it – see Figure 13.

The results of testing both described designs are presented below.

## 7.2

### *Evaluation module test*

The evaluation system was designed and tested in Polish. For the test, we used the “Tavern” story and the data gathered and checked in the study described in Section 5.





For the EM-A (see Figure 12) the following instruction was provided to the user.

Poniżej są cztery odpowiedzi na pytanie “Co niepokojącego ostatnio wydarzyło się w karczmie?”. Każdą z nich przyporządkuj do postaci i typu odpowiedzi. Tylko jedna odpowiedź pasuje do każdej z postaci, nie będą się powtarzać. Zaznacz też, jaką pewność, że Twoja odpowiedź jest poprawna. Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność.

*Below you will find four responses to the question: “What was the worrying thing that happened at the tavern?” Take each response and match them with the characters from the story and*

### Połącz odpowiedź z postacią

Poniżej są cztery odpowiedzi na pytanie "Co niepokojącego ostatnio wydarzyło się w karczmie?". Każdą z nich przyporządkuj do postaci i typu odpowiedzi. Tylko jedna odpowiedź pasuje do każdej z postaci, nie będą się powtarzać. Zaznacz też, jaką pewność, że Twoja odpowiedź jest poprawna. Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność.

Straż grodowa	Wspólnik	Pachołek	Żona
			
Twoja odpowiedź była bezpośrednia i zgodna z prawdą	Twoja odpowiedź była wymijająca, ale uprzejma.	Twoja odpowiedź była wymijająca i niezbyt uprzejma.	Twoja odpowiedź była prawdziwa, ale nie bezpośrednia.


  


Sam sprzątałeś przed chwilą wychodek, to sam sobie odpowiedz na to pytanie. Doniosłeś w końcu te beczki z piwem?


Ach, wiesz jak zwykle to bywa w tawernach – trochę szalonych ludzi, trochę pijanych, lecz raczej nic groźnego. Zresztą, przecież pracujesz tu nie od wczoraj.

W karczmie pojawił się tajemniczy mężczyzna, który chciał za zapłatą zostawić mi do przechowania pakunek z nieznaną zawartością, który miałbym następnie przekazać dalej.

Był jeden typ, ale nie obawiaj się.

 Pachołek
   
 Twoja pewność - 3 +

 Wybierz postać
   
 Twoja pewność - 1 +

 Wybierz postać
   
 Twoja pewność - 1 +


 Wybierz postać
   
 Twoja pewność - 1 +

Figure 12: Evaluation system A – description in the text

Figure 13:  
Evaluation  
system B –  
description in  
the text

**Jaka jest ta odpowiedź?**

Był tu ostatnio podejrzanym człowiekiem... Jak on wyglądał?

Był wysoki, miał długą ciemną szatę i kaptur na głowie. Miał dość bladą twarz i długie rude włosy.

**Ta odpowiedź jest:**

- Bezpośrednia i zgodna z prawdą
- Wymijająca, ale uprzejma
- Wymijająca i nieuprzejma
- Pośrednia i zgodna z prawdą

**Na ile masz pewność, że odpowiedź trafi do dobrej kategorii?**  
Odpowiedz w skali od 1 do 3. Im wyższa liczba, tym większa pewność.

- 1 +

*the answer type. Each character can be matched with only one response. Choose how certain you are in regard to the correctness of your answer. Answer on the scale of: 1 to 3. The more certain you are, the higher your answer should be.*

Four matching tasks were prepared for the EM-A. The data retrieved from the study described in Section 5 used for this module is presented in Table 13. For each question, a user is presented with four different responses. Each time, the order of response types is different. Table 13 presents this order.

For the second annotation module design, EM-B, the instruction was simply: What kind of response is that? The data used for the EM-B is presented in Table 14. Analogously to design A, here four tasks were prepared.

### 7.3

#### Procedure

The user study was conducted with the use of a dedicated website, and the answers were gathered online. Before starting the study, the users

Table 13: The data used for the first evaluation module EM-A

The original version	English translation
<p><b>Q1: Co niepokojącego ostatnio wydarzyło się w karczmie?</b></p> <p>(DA) Odwiedził mnie tajemniczy człowiek.</p> <p>(EAI) Nic. Przecież cały dzień tu siedzisz, to ciebie powinniśmy zapytać.</p> <p>(EAP) Jak to w karczmie, codziennie jakieś przygody.</p> <p>(IA) Nic, czym z czym już sobie nie poradziłem, był tu taki jeden</p>	<p><b>Q1: What was the worrying thing that happened at the tavern?</b></p> <p>I was visited by a mysterious man.</p> <p>Nothing. You sit here all day, we should be asking you.</p> <p>As usual in the tavern, new adventures every day.</p> <p>Nothing I couldn't deal with, some guy stopped by</p>
<p><b>Q2: Był tu ostatnio podejrzany człowiek. Jak on wyglądał?</b></p> <p>(EAI) a czy Ja muszę wszystkich pamiętać.</p> <p>(EAP) Był typowym wędrowcem, niczym się nie wyróżniał</p> <p>(IA) Trochę jak Twój kuzyn, Edmund, tylko wyższy.</p> <p>(DA) Nie widziałem wiele z powodu kaptura ale miał dość bladą cerę i rude włosy oraz był bardzo wysoki.</p>	<p><b>Q2: A suspicious man came by recently. What did he look like?</b></p> <p>is it Me who has to remember everyone.</p> <p>He was a typical vagabond, there was nothing special about him</p> <p>A little like your cousin, Edmund, just taller</p> <p>I couldn't see much because of his hood but he was quite pale, red haired, and very tall.</p>
<p><b>Q3: Co od niego dostałeś i jak to wyglądało?</b></p> <p>(EAP) Zamknięta, nie wiadomo co w środku, ale to taka przysługa tylko, powinniśmy być mili dla klientów jeśli chcemy mieć większy utarg.</p> <p>(EAI) Nie wiem o co ci chodzi. Zajmij się swoją pracą</p> <p>(DA) Dużą paczkę przewiązana lnianym sznurem</p> <p>(IA) Wyglądało jak pościel pod łóżkiem w naszym pokoju</p>	<p><b>Q3: What did you get from him and how did it look?</b></p> <p>It was closed, hard to tell what was inside, but it was just a favour we should be nice to clients if we want to have a bigger turnover</p> <p>I don't know what you're talking about. Get back to work</p> <p>A big package with a linen ribbon</p> <p>It looked like the sheets we keep under the bed in our room</p>
<p><b>Q4: Co ci za to zaoferował?</b></p> <p>(DA) 3 dukaty.</p> <p>(IA) tak ze sześć razy tyle, co nasze całe wesele nas wyszło, a skromne to było wesele, a skromne (mruga okiem).</p> <p>(EAI) Co mi zaoferował, to mi zaoferował.</p> <p>(EAP) Nie wspominał konkretnie</p>	<p><b>Q4: What did he offer you?</b></p> <p>3 ducats.</p> <p>about six times as much as we paid for our wedding reception, and it was a modest one, definitely modest (winks).</p> <p>What he offered me, he offered me.</p> <p>He didn't mention anything specific</p>

Table 14: Q-R pairs for the second annotation mode EM-B

The original version	English translation
<b>Q1: Co niepokojącego ostatnio wydarzyło się w karczmie?</b> Pojawiło się kilku podejrzanych typów, ale to nic szczególnego (EAP)	<b>Q1: What was the worrying thing that happened at the tavern?</b> A few suspicious guys came here, but it's nothing out of the ordinary.
<b>Q2: Był tu ostatnio podejrzany człowiek. Jak on wyglądał?</b> Trochę jak Twój kuzyn, Edmund, tylko wyższy. (IA)	<b>Q2: A suspicious man came by recently. What did he look like?</b> A little like your cousin, Edmund, just taller
<b>Q3: Co od niego dostałeś i jak to wyglądało?</b> Po co ci te wszystkie informacje? Szpiegujesz nas? (EAI)	<b>Q3: What did you get from him and how did it look?</b> Why do you need all this information? Are you spying on us?
<b>Q4: Co ci za to zaoferował?</b> 3 złote dukaty (DA)	<b>Q4: What did he offer you?</b> 3 golden ducats

could read information about it, then they had to agree to take part. Having done that, users were presented with a series of eight tasks in the two evaluation modes. The structure of the study was the following. First, the introduction and instructions were displayed. Once the user had expressed their agreement, the story “The Tavern” was introduced. This was followed by four tasks in EM-A and afterward by four tasks in EM-B. At the end, users provided elementary demographic data.

## 7.4

*Study group*

32 participants took part in the study, aged 29–70 (mean = 28.03, SD = 11.27). 26 participants were female, 5 were male.

## 7.5

*Results*

The results were gathered on April 4–8, 2021. User solutions were compared with the predefined answers (see Tables 15 and 16) to establish the evaluation correctness.

Question / Correct response	Correctness (%)	Average certainty	SD for average certainty
Q1 / DA	72	2.47	0.80
Q1 / EAI	72	2.50	0.76
Q1 / EAP	63	2.34	0.83
Q1 / IA	41	2.09	0.86
Q2 / EAI	81	2.53	0.76
Q2 / EAP	75	2.16	0.85
Q2 / IA	78	2.03	0.86
Q2 / DA	90	2.44	0.84
Q3 / EAP	81	2.28	0.81
Q3 / EAI	78	2.44	0.84
Q3 / DA	81	2.41	0.87
Q3 / IA	90	2.50	0.80
Q4 / DA	81	2.56	0.80
Q4 / IA	94	2.53	0.72
Q4 / EAI	81	2.34	0.90
Q4 / EAP	72	2.19	0.93

Table 15:  
The correctness of responses provided by users of the evaluation module EM-A

For the evaluation module EM-A users were requested to perform 16 matchings of responses to characters who would provide these responses (i.e. to one of four response types). The lowest correctness in this task was IA response in the first question (41%) and the highest was also for IA but for the fourth question (94%). Detailed results are presented in Table 15.

A closer look at the data correctness presented in Table 15 suggests that a form of training example or a training session would be needed for the evaluation module. The somewhat surprising lowest and highest correctness percentage for IA may be better understood in light of an overall low correctness percentage for the first question presented to users.

As for EM-B, the correctness of users’ identification of responses types is presented in Table 16. Here, we also observe that the highest correctness level was observed for the IA responses. This needs further exploration as it indicates interesting user behaviour – IA is the most

Table 16:  
The correctness  
of responses  
provided  
by users  
of the evaluation  
module EM-B

Question / Correct response	Correctness (%)	Average certainty	SD for average certainty
Q1 / EAP	97	2.53	0.84
Q1 / IA	100	2.62	0.75
Q1 / EAI	72	2.00	0.72
Q1 / DA	67	2.03	0.86

difficult response type to generate but the easiest to identify (as this preliminary data suggest).

The overall inter-annotator ( $N=32$ ) agreement established with the Fleiss Kappa measure (with the use of R programming language and irr library) was slightly higher for EM-B (0.666) than for the EM-A (0.503).

We believe that the correctness of evaluations provided by the users is satisfactory. The proposed designs naturally need further study. The correctness may be further improved by implementing training mechanisms known from the scientific discovery games, like the aforementioned Galaxy Zoo (Lintott *et al.* 2008). Training examples should be presented before the target responses and provide instant feedback for the user.

The evaluation module design presented in this section offer a promising addition to QRGs. It is worth stressing that a researcher may still rely completely on the manual check of the data performed by expert(s). We can imagine different QRGs usage scenarios which depend on the main purpose of the gathered linguistic data.

## 8

## QRGS DATA AS A PART OF THE EROTETIC REASONING CORPUS

We decided to publish part of the data gathered during our QRGs evaluation studies. As a platform to do this, we decided to use the Erotetic Reasoning Corpus (ERC; Łupkowski *et al.* 2017).

ERC is a data set for research on natural question processing. The basic intuition is that we are dealing with question processing in a situation when a question is not followed by an answer but with a new



question or a strategy of reducing it into auxiliary questions. Usually, such a situation takes place when an agent wants to solve a certain problem (expressed in a form of an initial question) but is not able to reach a solution using his/her own information resources. Thus, new data, collected *via* questioning is necessary. The corpus consists of the language data collected in previous studies on the question processing phenomenon. The outcomes of three research projects are employed here. These are: Erotetic Reasoning Test (Urbański *et al.* 2016a), Quest-Gen (Ignaszak and Łupkowski 2017) and Mind Maze (Urbański *et al.* 2016b). All the data are in Polish, but the tagging schema is in English to make it more universal to use.

The tagging schema for the ERC has three layers:

1. *Structural* – representing the structure of tasks used for the aforementioned studies. Here we distinguish elements like: instructions, justifications, different types of questions and declaratives.
2. *Inferential* – which allows for recognising normative elements related to the logic of questions used.
3. *Pragmatic* – representing various events that may occur in the dialogue, like e.g. long pauses. It also contains tags that allow expression of certain events related to the types of tasks used (like e.g. when a forbidden question is used).

### QRGS data preparation

8.1

The data to be added to ERC were retrieved in the study described in Section 5 (the study in Polish checking textual vs graphical QRGS version).

The data generated by the first 20 participants was used. Each participant provided responses to all four scenarios of “The Tavern” story (see the whole story in Appendix B):

1. Guard: DA
2. Business partner: EAP
3. Minion: EAI
4. Wife: IA.

Each solution was saved into a separate file. Overall, we have 80 files (20 per scenario), with 17,426 words. Each file started with “The Tavern” story followed with the paragraph introduction a scenario. Then, we have questions and user-generated responses formatted in a dialogue-like fashion.

These 80 files were manually annotated with the appropriately modified and extended ERC tagset.

## 8.2

### *ERC tagset extensions and modifications*

When it comes to the structural layer, the QUESTION tag has been extended with the AQ-ANSWER to cover cases where a question is responded with a question.

An example is presented below:

P: Co od niego dostałeś i jak to wyglądało? / *What did you get from him and what did it look like?*

K: Kogo masz na myśli? / *Who do you mean?*

Pachołek: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego dostałeś i jak to wyglądało?</QUESTION>

Karczmarsz: <QUESTION A1="AQ-ANSWER" A4="3"><EAI>Kogo masz na myśli?</EAI></QUESTION>

Query-response “Kogo masz na myśli / *Who do you mean?*” is identified with the tag AQ-ANSWER. The attribute of A4 links question and response in a given data file. (Arguments of A4 are the consecutive numbers of question-response pairs in a given file, see Figure 14.)

The pragmatic layer received one extension and one new tag, which is required to address the type of data from QRGS. The already existing tag KEY-INFO was extended with the attributes characteristic to the story, i.e. *character*, *package* and *payment*. This allows for identification of key information appearing in the story and user-generated responses.

S: Co od niego dostałeś i jak to wyglądało? / *What did you get from him and how did it look?*

K: Czarny pakunek, szczelnie zamknięty. / *Black package, it was tightly wrapped.*

```
Pachołek: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego
dostałeś i jak to wyglądało?</QUESTION>
Karczmaz: <DECLARATIVE A1="AQ-ANSWER" A4="3"><DA><KEY-INFO A1="package"
A4="2">Czarny pakunek, szczelnie zamknięty.</KEY-INFO></DA></DECLARATIVE>
```

The response: “Czarny pakunek, szczelnie zamknięty. / *Black package, it was tightly wrapped*” is identified as a declarative answer to the question above and also as a key-info from the point of view of the story plot.

Another additional tag is RRT (required response type) with the attributes related to four response types generated by QRGS users: DA, IA, EAP, EAI and OTHER (for possible responses not fitting the expected categories).

80 QRGS files were annotated by two annotators with the updated ERC tagset. A sample annotated file is presented in Figure 14.

The annotated files were all checked in accordance with the procedure for the ERC described in Łupkowski *et al.* 2017. Firstly, all files were checked for the syntactic correctness of the XML tags with the Emacs editor (version 26.3) and Vacuous XML schema<sup>2</sup>. All identified errors were eliminated. In the next step, 50 files were chosen randomly and intra- and inter-annotator studies were performed. Kappa values were established with the use of the R programming language (version 3.5.0) and irr package. Results were satisfactory, as Cohen’s kappa for the intra-annotator study was 0.819 (with 84% agreement) and for the inter-annotation study was 0.791 (with 82% agreement).

The annotated QRGS data are now available as a part of the Erotetic Reasoning Corpus.<sup>3</sup>

---

<sup>2</sup><https://www.w3.org/TR/xmlschema11-1/>.

<sup>3</sup>Erotetic Reasoning Corpus homepage is: <https://ercorpus.wordpress.com/>. The latest version of ERC is available there along with documentation describing the tag-set used, and ERC tools: Search & Browse Tool (for browsing ERC files with and without annotation visible, as well as searching for particular ERC tags); XML/L<sup>A</sup>T<sub>E</sub>X Parser (easy transformation of XML files into L<sup>A</sup>T<sub>E</sub>X files); and ERC XML Schema (which allows for validating the annotation of ERC files).

```

1 <KORPUS A1="QRGS" A2="straz12">
2
3 <INSTRUCTION>
4 Historia pewnej karczmy: Straż grodowa
5 Jesteś właścicielem tawerny na obrzeżach miasta. Wczoraj zjawił się u ciebie <
6 KEY-INFO A1="character" A4="1"> tajemniczy podróżny, odziany w szeroki, ciemny
7 płaszcz z kapturem, który prawie całkiem zasłaniał jego twarz. Był bardzo wysoki
8 i dało się zaobserwować niezwykle bladą twarz i wystające spod kaptura długie,
9 rude włosy.</KEY-INFO> W ręku trzymał <KEY-INFO A1="package" A4="2">niemały
10 pakunek, przewiązany lnianą wstęgą.</KEY-INFO> Położył go na kontuarze i
11 zaproponował ci układ. W zamian za <KEY-INFO A1="payment" A4="3">sowitą zapłatę
12 w postaci trzech dukatów</KEY-INFO>, tajemnicza przesyłka zostaje pod twoim
13 nadzorem. Ma zostać nietknięta, do czasu kiedy zjawi się człowiek, który będzie
14 znał ustalone hasło.
15
16 Następnego dnia do twojej tawerny wkracza straż grodowa. Wygląda na to, że dziś
17 nie przyszli na ciepły posiłek po służbie. Chcą zadać ci kilka pytań.
18 <RRT A1="DA">Chociaż pytają o tajemniczego wędrowca, uznajesz że mądrze będzie
19 odpowiadać im bezpośrednio i zgodnie z prawdą. Nie chcesz przecież popaść w
20 konflikt z władzą.</RRT>
21 </INSTRUCTION>
22
23 Straż: Witaj karczmarzu! Dziś przybывamy w sprawie służbowej. Mamy kilka pytań.
24
25 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="1">Co niepokojącego ostatnio
26 wydarzyło się w karczmie?</QUESTION>
27
28 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="1"><EAP>Nic szczególnego. Kilku
29 podpitych gości wszczęło bójki, ale to w zasadzie norma. </EAP></DECLARATIVE>
30
31 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="2">Był tu ostatnio podejrzany
32 człowiek. Jak on wyglądał?</QUESTION>
33
34 Karczmarz: <QUESTION A1="AQ-ANSWER" A4="2"><DA>Podejrzany człowiek?</DA></QUESTION>
35 <DECLARATIVE A1="AQ-ANSWER" A4="2"><DA> A no był taki jakiś dziwak.<KEY-INFO A1="
36 character" A4="1"> Chudy, blade, rude włosy. Nie widziałem twarzy, bo zakrywał ją
37 kaptur.</KEY-INFO></DA></DECLARATIVE>
38
39 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="3">Co od niego dostałeś i jak to
40 wyglądało?</QUESTION>
41
42 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="3"><DA><KEY-INFO A1="package" A4="2">
43 Dał mi jakiś spory pakunek na przechowanie dla kogoś innego.</KEY-INFO></DA></
44 DECLARATIVE>
45
46 Straż: <QUESTION A1="AUXILIARY" A3="OTHER" A4="4">Co ci za to zaoferował?</QUESTION
47 >
48
49 Karczmarz: <DECLARATIVE A1="AQ-ANSWER" A4="4"><DA><KEY-INFO A1="payment" A4="3">
50 Dał mi trzy dukaty. A to sporo za jakiś tam pakunek.</KEY-INFO></DA></DECLARATIVE>
51
52 </KORPUS>

```

Figure 14: An exemplary QRGs file annotated with the ERC tagset

This paper presents the concept of the Question Responses Generation System, a crowdsourced framework for gathering linguistic data of a specific form. QRGS allows for relatively simple and efficient retrieval of various responses to questions.

QRGS requires a simple story and follow up scenarios to the main plot which lead a user to provide responses of the required type. As such, it is a very universal framework. The stories are relatively simple and easy to write. The whole schema of the framework is also simple and – crucially – easy to implement. One does not need any special programming skills. As presented in the paper, even Google Forms (or any other similar platform) is enough to implement QRGS and gather data.

We presented a series of evaluation studies of QRGS. Seven stories in total were tested so far (and are available as appendices for this paper). Four of them are in Polish, three in English.

During our evaluation studies QRGS appeared to be effective in terms of the amount of data gathered.<sup>4</sup> Altogether, 4,304 responses to questions have been generated for Polish and 296 Q-R pairs have been generated for English. Also the correctness of the data is satisfactory, as summarised in Table 17. Correctness is understood here as compliance with the type of the response expected from a given story scenario. Our findings indicate that the most unproblematic response type in an unsupervised crowdsourced data generation are direct answers (DA). The most difficult for QRGS users are indirect answers (IA).

As the reported results show, the correctness level varies between studies and does not reach 100%. This indicates that the data gathered *via* QRGS cannot be straightforwardly used for certain applications, e.g., training data for language models. Such data needs to be evaluated first. That is why we also propose a promising and effective crowdsourcing solution that allows for data evaluation. Using one

---

<sup>4</sup>However, as pointed by the anonymous reviewer, the amount of data gathered may be dependent on many parameters, not only the framework supporting the acquisition, such as: availability of participants or the interval of time allocated for the crowdsourcing activity.

Table 17:  
The summary  
of the  
correctness  
of the data  
gathered with  
the use of QRGS

Study	Correctness (%)
Pilot (Section 3)	77
Native vs non-native (Section 4)	63
Textual vs graphical (Section 5)	49
Casual vs detective-like story (Section 6)	54

(or two) proposed evaluation modules for additional data correctness checks. Naturally, an expert manual check of the data is still possible (and recommended for certain future applications). After the evaluation phase we envisage two potential scenarios: 1) eliminating non-correct responses (as the relative cost of generating data with QRGS is not high, we see this as an acceptable option); 2) reusing non-correct responses for which correct labels are added during the evaluation stage (this leaves a researcher with the complete generated dataset).

In line with scenario 2, part of QRGS generated data was formatted, manually annotated, thoroughly checked and incorporated into the Erotetic Reasoning Corpus and is now publicly available.

The series of QRGS studies resulted also in several findings useful for future QRGS development and implementations.

1. No difference between native and non-native English speakers for correctness and the response length were observed. At least for English, we may expect valuable data as long as we gather users with good knowledge of the language. Naturally, this observation needs further study for other languages.
2. Very small differences were observed *vis a vis* correctness for the graphical vs text and casual vs detective-like stories; similarly, no difference between the text condition and the graphical condition for the response length. This suggests that the simple, textual version is enough to effectively use QRGS.
3. Participants in the text condition were more engaged in the task (than in the graphical condition). This is an interesting and somewhat surprising finding suggesting (as in the case of 2) that the text-only version of QRGS may be a better solution.
4. Observed differences in the response length for the casual vs detective-like stories. This effect suggests that the detective-like stories may result in more extended responses. This needs further

study – especially quantitatively, where users’ experiences would be evaluated.

QRGS offers a promising framework for gathering large amounts of various types of responses to questions. We believe that it needs further testing with other languages, especially those which have lower spoken language corpora coverage. There are also open questions which may be addressed when it comes to the QRGS idea, e.g. how to increase the data correctness level, especially for IA? Or how to add more scenarios to the stories, such that more response categories would be generated (and the whole QRGS task would not get boring and time-consuming for users)?

#### ACKNOWLEDGEMENTS

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the program “Investissements d’Avenir” (reference: ANR-10-LABX-0083). It contributes to the IdEx Université de Paris – ANR-18-IDEX-0001. Work on the Erotetic Reasoning Corpus was supported by the National Science Centre, Poland (DEC-2013/10/E/HS1/00172 and DEC-2012/04/A/HS1/00715).

We would also like to give our thanks to three anonymous reviewers for the *Journal of Language Modelling* for their insightful comments on this article.

## APPENDICES

### A STORIES FOR THE NATIVE / NON-NATIVE STUDY

**Story I. EPILEPSY.** Your co-worker Anna just had an epilepsy attack. You are aware this happens sometimes, as for the safety reasons she has informed you some time ago. You also know it has been 6 months since her last seizure event. Today was just an ordinary day and nothing uncommon preceded the attack. When she lost consciousness and fell to the floor you were standing next to her. You have assisted Anna making sure she does not hurt herself during the convulsions. You measured the length of the attack – it took about 5 minutes. After that, she did not regain consciousness, so you have decided to call for emergency.

**Scenario A.** The paramedic has arrived and is asking you some questions about Anna and you want your answers to be very accurate:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

**Scenario B.** Anna's mother is calling you because she could not reach her daughter. You told her about the attack and now she has more questions. Unfortunately, you are still in the office and you do not want people around you to overhear the details about Anna. As you cannot leave the common space and your colleagues suspect the topic of the conversation, you will need to answer indirectly:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

**Scenario C.** Matt from your team came by as he heard Anna had been taken to the hospital. He seems worried and is asking you questions



about Anna's condition. From what you know he and Anna are friends but Anna emphasised that she shared the information in secret, so you feel obliged to keep it. You understand his concerns but you are not going to reveal anything without permission:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

**Scenario D.** Rob, the annoying colleague from another department came by. He is known for his terrible gossiping habit and now is asking you questions about Anna's condition and information she told you in confidence. His behaviour irritates you and you do not want to talk with him about Anna. How will you react to his questions?:

1. How long did the seizure last?
2. Was she conscious during the attack?
3. Did anything unusual happen before the accident?
4. How many attacks did she have lately?

The paramedic team was able to rouse Anna and she seems all good but will be taken to the hospital for observation.

**Story II. SECRET SANTA.** You and your friends have decided to organise a Secret Santa event this year. Each member of your pack will receive a gift prepared jointly by the rest of the group members. After a short brainstorming session you proposed to give Joe the Craft Beer Brewing Kit and this idea was met with great enthusiasm. It costs 50 USD and this sum will be divided evenly between 5 people. You are responsible for collecting the money and purchasing the kit. You are going to make the purchase on Friday so your friends should give you their shares until then.

**Scenario A.** George (one of the conspiracy group) is not familiar with the arrangements and has just visited you for details. You can speak openly with George about the organisational details:

1. What are we giving to Joe?
2. How much is the contribution rate?

3. Who will make the purchase?
4. What is the deadline for collecting the money?

**Scenario B.** Jane was not present at the brainstorm meeting. She has called you and has some questions but Joe is in a car with you. You want to pass the information to Jane while hiding it from Joe. Try to provide indirect answers to the following questions:

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

**Scenario C.** Maggie, Joe's sister, is wishing to participate, too. You like her, but you do not trust her. She might share the secrets with Joe. She wants to know the details. Try to decline her in a polite manner:

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

**Scenario D.** Joe is extremely sneaky and is trying to draw some information on his gift from you. He has sent his younger brother to spy on you. You want to teach him a lesson of minding his own business and decline him in a rather rude way. How will you react to his questions?

1. What are we giving to Joe?
2. How much is the contribution rate?
3. Who will make the purchase?
4. What is the deadline for collecting the money?

## **B**                      **STORY FOR THE GRAPHICAL / TEXTUAL STUDY**

**Story: TAWERNA / TAVERN.** Jesteś właścicielem tawerny na obrzeżach miasta. Wczoraj zjawił się u ciebie tajemniczy podróżny, odziany

w szeroki, ciemny płaszcz z kapturem, który prawie całkiem zasłaniał jego twarz. Był bardzo wysoki i dało się zaobserwować niezwykle bladą twarz i wystające spod kaptura długie, rude włosy. W rękę trzymał niemały pakunek, przewiązany lnianą wstęgą. Położył go na kontuarze i zaproponował ci układ. W zamian za sówitą zapłatę w postaci trzech dukatów, tajemnicza przesyłka zostaje pod twoim nadzorem. Ma zostać nietknięta, do czasu kiedy zjawi się człowiek, który będzie znał ustalone hasło.

*You are the owner of a tavern in the suburbs. Yesterday a mysterious stranger came to you. He was wearing a wide, dark coat with a hood which covered almost all of his face. The stranger was very tall and he was extremely pale. You could observe long ginger hair under his hood. In his hand, he carried a significantly sized package with a linen ribbon. He put it on the counter and offered you a deal. He wanted to give the package to you for safekeeping and in turn he would pay you a fair price of 3 ducats. The package is to be left untouched until a man comes and tells you a password.*

**Scenario GUARD.** Następnego dnia do twojej tawerny wkracza straż grodowa. Wygląda na to, że dziś nie przyszli na ciepły posiłek po służbie. Chcą zadać ci kilka pytań. Chociaż pytają o tajemniczego wędrowca, uznajesz że mądrze będzie odpowiadać im bezpośrednio i zgodnie z prawdą. Nie chcesz przecież popaść w konflikt z władzą. (Straż grodowa) Witaj karczmarzu! Dziś przybywamy w sprawie służbowej. Mamy kilka pytań.

*On the next day, guards come to your tavern. It seems like they're not here to eat something warm after work. They want to ask you a few questions. Even though they're asking about the mysterious stranger, you decide that it will be wise to reply to them directly and truthfully. You don't want to get into a conflict with the guards. (Guards) Hello, innkeeper! Today we're here on business. We have a few questions to ask you.*

**Scenario BUSINESS PARTNER.** Wieczorem zaczyna cię twój wspólnik, którego wczoraj nie było w gospodzie. Słyszał plotki od innych pracowników, dlatego postanawia wypytać cię o szczegóły. Odpowiedz mu wymijająco, ale uprzejmie – w końcu to twój wspólnik. (Wspólnik) Cześć chłopie! Dawno cię nie widziałem. Mam nadzieję, że wczorajszy obrót był wysoki. Muszę się przyznać, że słyszałem niepokojące plotki.

*In the evening, your business partner comes up to you. He wasn't in the tavern yesterday. He's heard some gossip from other employees and he wants to know more details. Answer him in an evasive, but polite way, he's your business partner after all.*

*(Business partner) Hi, man! I haven't seen you in a while. I hope that yesterday's turnover was high. If I'm being honest, I've heard some unnerving rumours.*

**Scenario MINION.** Tego samego wieczoru podchodzi do ciebie jeden ze sług zatrudnionych w karczmie. On również słyszał plotki. Sam dobrze wiesz, że te lubią rozchodzić się w zastraszającym tempie. Pachołek ma do ciebie parę pytań. Odpowiedz mu wymijająco – nie musisz być dla niego szczególnie uprzejmy. (Pachołek) Panie, wiem że ja tu tylko sprzątam, ale chciałbym cię o coś zapytać.

*The same night one of the minions who work at your tavern comes to you. He's also heard the gossip. You know how fast they spread. The minion has a couple of questions for you. Answer him in an evasive way – you don't need to be polite. (Minion) Good sir, I know I'm a simple cleaner, but I would like to ask you about something.*

**Scenario WIFE.** Kolejnego dnia z rana żona również bierze cię na wypytki. Ponieważ rozmowa toczy się przy kontuarze, przysłuchują się jej jak zawsze zaciekawieni goście gospody. Postaraj się udzielić żonie prawdziwych informacji, ale nie w bezpośredni sposób. (Żona) Witaj mężu. Mam nadzieję, że dobrze spałeś. Dopiero dziewiąta rano, a goście już pytają o zupę. Słyszałam od współnika niepokojące informacje – podobno odwiedziła nas straż grodowa. Mógłbyś mi rozjaśnić sprawę.

*Next morning your wife wants to have a chat with you. The conversation is taking place at the counter, so as usual, curious tavern guests are listening out for information. Try to give your wife true information, but do it indirectly. (Wife) Hello, husband. I hope you slept well. It's only 9 in the morning and the guests are already asking for soup! Your business partner has given me some worrisome information – apparently the guards visited us. You could tell me what happened.*

## QUESTIONS

1. Co niepokojącego ostatnio wydarzyło się w karczmie? / *What was the worrying thing that happened at the tavern?*

2. Był tu ostatnio podejrzany człowiek. Jak on wyglądał? / *A suspicious man came by recently. What did he look like?*
3. Co od niego dostałeś i jak to wyglądało? / *What did you get from him and how did it look?*
4. Co ci za to zaoferował? / *What did he offer you?*

## STORIES FOR THE PLOT FORMULATION STUDY

C

**Story I. KRADZIEŻ BIŻUTERII / JEWELRY THEFT.** Z domu szanowanego hrabiego ukradziono cenną, rodzową biżuterię. Jako zaufany kucharz, tej nocy przygotowywałeś dla pana domu kolację i przypadkowo wpadłeś na złodzieja, któremu jednak udało się uciec. Zdążyłeś mu się przyjrzeć, ale niestety nie widziałeś jego twarzy. Mimo to, wiesz, że: Złodziej był wysokim, szczupłym mężczyzną w ciemnej kurtce z kapturem. Biżuterię wyniósł w pudełku na buty. Złodziej uciekł z domu przez tylne wyjście. Kradzież nastąpiła około godziny 20.

*From the respected count's house, valuable ancestral jewellery was stolen. As a trusted chef, you were preparing dinner for the household that night and accidentally stumbled upon the thief, who managed to escape. You had a chance to observe them, but unfortunately did not see his face. Nonetheless, you know that: The thief was a tall, slim man wearing a dark jacket with a hood. He carried the jewelry out in a shoebox. The thief fled the house through the back exit. The theft occurred around 8 p.m.*

**Scenario SZEF POLICJI / POLICE DIRECTOR.** Na miejscu zjawia się szef policji, który próbuje ustalić szczegóły kradzieży. Powinieneś udzielić prawdziwych i precyzyjnych odpowiedzi na jego pytania.  
*When the chief of police arrives at the scene to investigate the details of the theft, you should provide true and precise answers to his questions.*

**Scenario OFIARA KRADZIEŻY / THEFT VICTIM.** Zostałeś poinstruowany, aby tymczasowo nie dzielić się szczegółami śledztwa z panem domu, ze względu na jego słabe zdrowie. Ponieważ jednak hrabia

próbuję wypytać Cię o zajście, postaraj się dać mu do zrozumienia, że nie możesz udzielić odpowiedzi, jednak zrób to w sposób uprzejmy (w końcu to Twój pracodawca).

*You've been instructed not to share details of the investigation with the count temporarily, due to his poor health. However, since the count is attempting to inquire about the incident, try to politely indicate that you cannot provide an answer, keeping in mind that he is your employer.*

**Scenario SŁUŻBA / MINIONS.** Zaraz po udaniu się hrabiego do sypialni, podchodzi do Ciebie kilka osób ze służby. Ze względu na trwające śledztwo nie możesz udzielić im bezpośrednio informacji, jednak postaraj się odpowiedzieć zgodnie z prawdą.

*Right after the count goes to his bedroom, a few members of the household staff approach you. Due to the ongoing investigation, you cannot directly provide them with information, but try to answer truthfully.*

**Scenario SĄSIAD / NEIGHBOUR.** Następnego dnia spotykasz sąsiada hrabiego, którego sylwetka, według Ciebie, łądząco przypomina złodzieja biżuterii (o czym wspomniałeś także policjantom). Odpowiedz na jego pytania w taki sposób, żeby zrozumiał, że nie chcesz z nim rozmawiać (nie musisz być bardzo uprzejmy, jednak nie powinieneś kłamać ani zbyć go słowami „nie wiem”, ponieważ nie możesz pozwolić, aby domyślił się, że jest podejrzanym).

*The next day, you encounter the count's neighbour, whose silhouette, in your opinion, strikingly resembles that of the jewellery thief (which you also mentioned to the police). Answer his questions in a way that makes him understand you don't want to engage in conversation (you don't have to be overly polite, but you shouldn't lie or brush him off with "I don't know," as you cannot allow him to suspect he's a suspect).*

## QUESTIONS

1. Jak wyglądał złodziej? / *What did the thief look like?*
2. W czym udało mu się wynieść biżuterię? / *How did he manage to carry the jewellery?*
3. W jaki sposób uciekł z domu? / *How did he escape from the house?*
4. O której godzinie zdarzyła się kradzież? / *At what time did the theft occur?*

**Story II. ZARĘCZYNY / ENGAGEMENT.** Twój dobry przyjaciel Piotr poprosił Cię o pomoc w organizacji imprezy-niespodzianki, na której chce oświadczyć się swojej dziewczynie Ewie. Twoim zadaniem jest potwierdzenie rezerwacji w restauracji oraz zadbanie o zaproszenie zaufanych gości. Przyjaciel zostawił Ci kilka wskazówek: Impreza ma zacząć się o godzinie 17 i potrwa do północy. Zaproszonych będzie 15 osób, w tym rodzice Piotra i Ewy. Na imprezie będzie podawane ulubione wino pary. W torcie przygotowanym na imprezę zostanie ukryty pierścionek zaręczynowy.

*Your good friend Piotr has asked you for help in organising a surprise party, where he plans to propose to his girlfriend Ewa. Your task is to confirm the restaurant reservation and ensure the invitation of trusted guests. Your friend left you some guidelines: The party is to start at 5 p.m. and last until midnight. 15 people will be invited, including Piotr and Ewa's parents. The couple's favorite wine will be served at the party. An engagement ring will be hidden in the cake prepared for the party.*

**Scenario SZEF RESTAURACJI / RESTAURANT MANAGER.** Na umówionym spotkaniu omawiasz szczegóły przyjęcia z szefem restauracji. Odpowiedz na jego pytania jak najdokładniej, aby wiedział, jak się przygotować na imprezę.

*At the scheduled meeting, you discuss the details of the reception with the restaurant manager. Answer his questions as accurately as possible so he knows how to prepare for the event.*

**Scenario TELEFON OD MACIEJA / PHONE FROM MACIEJ.** Po rozmowie z szefem restauracji jedziesz spotkać się z Ewą. Podczas Waszego spotkania dzwoni do Ciebie brat Piotra, Maciej, który wie o imprezie i chce dopytać Cię o szczegóły. Odpowiedz na jego pytania w zrozumiałym sposób, ale tak, aby Ewa nie domyśliła się, o czym rozmawiacie.

*After the conversation with the restaurant manager, you go to meet Ewa. During your meeting, Piotr's brother, Maciej, who knows about the party, calls you and wants to inquire about the details. Answer his questions in an understandable way, but ensure Ewa doesn't suspect what you're discussing.*

**Scenario EWA / EWA.** Po zakończonej rozmowie z Maciejem, Ewa próbuje wypytać Cię o to, o czym rozmawialiście. Wie ona tylko, że im-

preza odbędzie się w najbliższą sobotę, jednak cała reszta powinna pozostać niespodzianką. Postaraj się odpowiedzieć na jej pytania tak, aby dać do zrozumienia, że nie możesz jej nic zdradzić, ale bądź uprzejmy, aby jej nie zdenerwować.

*After the conversation with Maciej, Ewa tries to ask you about what you talked about. She only knows that the party will take place next Saturday, but everything else should remain a surprise. Try to answer her questions in a way that implies you can't reveal anything, but be polite so as not to upset her.*

**Scenarij ZNAJOMA / FRIEND.** Gdy wracasz do domu po spotkaniu, spotykasz na ulicy znajomą, za którą nie przepadają Ewa i Piotr. Nie jest ona zaproszona na imprezę, jednak usłyszała o niej od swojego kolegi. Odpowiedz na jej pytania tak, aby zrozumiała, że nie chcesz z nią rozmawiać (nie musisz być bardzo uprzejmy, jednak nie powinieneś kłamać ani zbyć jej słowami „nie wiem”).

*When you return home after the meeting, you meet an acquaintance on the street, whom Ewa and Piotr don't particularly like. She's not invited to the party, but she heard about it from her friend. Answer her questions in a way that makes her understand you don't want to talk to her (you don't have to be very polite, but you shouldn't lie or brush her off with "I don't know").*

## QUESTIONS

1. W jakich godzinach odbędzie się impreza? / *What time will the party take place?*
2. Ile osób jest na nią zaproszonych? / *How many people are invited?*
3. Jaki alkohol zostanie podany na imprezie? / *What alcohol will be served at the party?*
4. Jakie są zaplanowane niespodzianki? / *What surprises are planned?*



## REFERENCES

- Anne H. ANDERSON, Miles BADER, Ellen Gurman BARD, Elizabeth H. BOYLE, Gwyneth M. DOHERTY, Simon C. GARROD, Stephen D. ISARD, Jacqueline C. KOWTKO, Jan M. MCALLISTER, Jim MILLER, Catherine F. SOTILLO, Henry S. THOMPSON, and Regina WEINERT (1991), The HCRC Map Task Corpus, *Language and Speech*, 34(4):351–366.
- Lou BURNARD, editor (2007), *Reference guide for the British National Corpus (XML Edition)*, Oxford University Computing Services on behalf of the BNC Consortium, <http://www.natcorp.ox.ac.uk/XMLedition/URG/>, access 20.03.2017.
- Jon CHAMBERLAIN, Massimo POESIO, and Udo KRUSCHWITZ (2008), PhraseDetectives: A web-based collaborative annotation game, in *Proceedings of the International Conference on Semantic Systems (I-Semantics' 08)*, pp. 42–49.
- Seth COOPER, Adrien TREUILLE, Janos BARBERO, Andrew LEAVER-FAY, Kathleen TUIITE, Firas KHATIB, Alex Cho SNYDER, Michael BEENEN, David SALESIN, David BAKER, Zoran POPOVIĆ, and > 57,000 Foldit PLAYERS (2010), The challenge of designing scientific discovery games, in *Proceedings of the Fifth International Conference on the Foundations of Digital Games*, pp. 40–47.
- Cristian DANESCU-NICULESCU-MIZIL and Lillian LEE (2011), Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, in *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics*, pp. 76–87, Association for Computational Linguistics.
- Lorna DSILVA, Shubhi MITTAL, Brian KOEPNICK, Jeff FLATTEN, Seth COOPER, and Scott HOROWITZ (2019), Creating custom foldit puzzles for teaching biochemistry, *Biochemistry and Molecular Biology Education*, 47(2):133–139.
- Dagmara DZIEDZIC (2016), Use of the free to play model in games with a purpose: the RoboCorp game case study, *Bio-Algorithms and Med-Systems*, 12(4):187–197.
- Jonathan GINZBURG, Zulipiye YUSUPUJIANG, Chuyuan LI, Kexin REN, Aleksandra KUCHARSKA, and Paweł LUPKOWSKI (2022), Characterizing the response space of questions: data and theory, *Dialogue & Discourse*, 13(2):79–132.
- Jonathan GINZBURG, Zulipiye YUSUPUJIANG, Chuyuan LI, Kexin REN, and Paweł LUPKOWSKI (2019), Characterizing the response space of questions: a corpus study for English and Polish, in *Proceedings of the 20th annual SIGdial meeting on discourse and dialogue*, pp. 320–330.

Oliwia IGNASZAK and Paweł ŁUPKOWSKI (2017), Inferential Erotetic Logic in modelling of cooperative problem solving involving questions in the QuestGen game, *Organon F*, 24(2):214–244.

Charlene JENNETT, Anna L. COX, Paul CAIRNS, Samira DHOPAREE, Andrew EPPS, Tim TIJS, and Alison WALTON (2008), Measuring and defining the experience of immersion in games, *International Journal of Human-Computer Studies*, 66(9):641–661.

Chris J. LINTOTT, Kevin SCHAWINSKI, Anže SLOSAR, Kate LAND, Steven BAMFORD, Daniel THOMAS, M. Jordan RADDICK, Robert C. NICHOL, Alex SZALAY, Dan ANDREESCU, Phil MURRAY, and Jan VANDENBERG (2008), Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey, *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Paweł ŁUPKOWSKI and Dagmara DZIEDZIC (2016), Building players' engagement – a case study of games with a purpose in science, *Homo Ludens*, 1(9):127–145.

Paweł ŁUPKOWSKI, Mariusz URBAŃSKI, Andrzej WIŚNIEWSKI, Wojciech BŁĄDEK, Agata JUSKA, Anna KOSTRZEWA, Dominika PANKOW, Katarzyna PALUSZKIEWICZ, Oliwia IGNASZAK, Joanna URBAŃSKA, Natalia ŻYLUK, Andrzej GAJDA, and Bartosz MARCINIAK (2017), Erotetic Reasoning Corpus. A data set for research on natural question processing, *Journal of Language Modelling*, 5(3):607–631.

Paweł ŁUPKOWSKI and Patrycja WIETRZYCKA (2015), Gamification for question processing research – the QuestGen game, *Homo Ludens*, 7(1):161–171.

Adam PRZEPIÓRKOWSKI, Mirosław BAŃKO, Rafał L. GÓRSKI, Barbara LEWANDOWSKA-TOMASZCZYK, Marek ŁAZIŃSKI, and Piotr PEŹIK (2011), National Corpus of Polish, in *Proceedings of the 5th language & technology conference: Human language technologies as a challenge for computer science and linguistics*, pp. 259–263, Fundacja Uniwersytetu im. Adama Mickiewicza Poznań.

Piotr PEŹIK (2014), Spokes search engine for Polish conversational data, <http://hdl.handle.net/11321/47>, CLARIN-PL digital repository.

Carolyn P. ROSÉ, Barbara Di EUGENIO, and Johanna D. MOORE (1999), A dialogue-based tutoring system for basic electricity and electronics, in Susanne P. LAJOIE and Martial VIVET, editors, *Artificial intelligence in education*, pp. 759–761, IOS, Amsterdam.

Paweł STROJNY and Agnieszka STROJNY (2014), Kwestionariusz immersji – polska adaptacja i empiryczna weryfikacja narzędzia, *Homo Ludens*, 1(6):171–186.

Mariusz URBAŃSKI, Katarzyna PALUSZKIEWICZ, and Joanna URBAŃSKA (2016a), Erotetic problem solving: From real data to formal models. An analysis

of solutions to erotetic reasoning test task, in F. PAGLIERI, L. BONETTI, and S. FELLETTI, editors, *The Psychology of Argument: Cognitive Approaches to Argumentation and Persuasion*, pp. 33–46, College Publications.

Mariusz URBAŃSKI, Natalia ŻYLUK, Katarzyna PALUSZKIEWICZ, and Joanna URBAŃSKA (2016b), A formal model of erotetic reasoning in solving somewhat ill-defined problems, in D. MOHAMMED and M. LEWIŃSKI, editors, *Argumentation and Reasoned Action. Proceedings of the 1st European Conference on Argumentation. London: College Publications*, pp. 973–983, College Publications.

Noortje VENHUIZEN, Valerio BASILE, Kilian EVANG, and Johan BOS (2013), Gamification for word sense labeling, in *Proceedings of the 10th International Conference on Computational Semantics (IWCS'13) – Short Papers*, pp. 397–403.

Anthony J. VIERA and Joanne M. GARRETT (2005), Understanding interobserver agreement: the kappa statistic, *Family Medicine*, 37(5):360–363.

Aleksandra WASIELEWSKA and Paweł ŁUPKOWSKI (2022), IMUW the questionnaire measuring the engagement of attention in a task execution, <https://osf.io/6dt8f/>.

Andrzej WIŚNIEWSKI (2013), *Questions, inferences and scenarios*, College Publications, London.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2020), Designing a GWAP for collecting naturally produced dialogues for low resourced languages, in *Workshop on Games and Natural Language Processing*, pp. 44–48.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2021), Data collection design for dialogue systems for low-resource languages, *Conversational Dialogue Systems for the Next Decade*, pp. 387–392.

Zulipiye YUSUPUJIANG and Jonathan GINZBURG (2022), UgChDial: A Uyghur chat-based dialogue corpus for response space classification, in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 3140–3149.

*Paweł Łupkowski*

Ⓘ 0000-0002-5335-2988

Pawel.Lupkowski@amu.edu.pl

*Ewelina Chmurska*

*Adrianna Płatosz*

*Aleksandra Kwiecień*

*Barbara Adamska*

*Magdalena Szkalej*

Faculty of Psychology and Cognitive  
Science

Adam Mickiewicz University

Szamarzewskiego 89a, 60-568 Poznań

*Jonathan Ginzburg*

Ⓘ 0000-0001-5737-0991

yonatan.ginzburg@u-paris.fr

Université Paris Cité, CNRS,  
Laboratoire de Linguistique Formelle  
5 Rue Thomas Mann,  
75205, Paris

Paweł Łupkowski, Jonathan Ginzburg, Ewelina Chmurska, Adrianna Płatosz, Aleksandra Kwiecień, Barbara Adamska, and Magdalena Szkalej (2024), *QRGS – Question Responses Generation via crowdsourcing*, *Journal of Language Modelling*, 12(1):213–270

Ⓙ <https://dx.doi.org/10.15398/jlm.v12i1.372>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

Ⓒ <http://creativecommons.org/licenses/by/4.0/>