# What Did This Castle Look Like Before? Exploring Referential Relations in Naturally Occurring Multimodal Texts

**Ronja Utescher**
Friedrich-Schiller-University Jena/
Bielefeld University
r.utescher@uni-bielefeld.de

**Sina Zarrieß**
Bielefeld University
sina.zarriess@uni-bielefeld.de

## Abstract

Multi-modal texts are abundant and diverse in structure, yet Vision & Language research of these naturally occurring texts has mostly focused on genres that are comparatively light on text, like tweets. In this paper, we discuss the challenges and potential benefits of a V&L framework that explicitly models referential relations, taking Wikipedia articles about buildings as an example. We briefly survey existing related tasks in V&L and propose multi-modal information extraction as a general direction for future research.

## 1 Introduction

Many types of naturally occuring texts are inherently multi-modal: articles, social media posts, recipes, encyclopedias, manuals, advertisement, comics, etc. Research on semiotics has long noted that the relationship between the linguistic and visual elements of such texts is extremely complex (Hardy-Vallée, 2016) and varies widely across genres (Delin and Bateman, 2002). To date, research in Vision & Language, however, has mostly focussed on crowdsourced data that simply aligns relatively short snippets of text to images (e.g. Wu et al. (2017)), sequences of images (e.g. Yang et al. (2019)) or video (e.g. Pan et al. (2020)). Here, the text-image relationship is simplified to a substantial, if not artificial, degree.

In this paper, we take a qualitative look at some examples of real-world multi-modal texts, i.e. Wikipedia articles on entities of the type "building". We find that many phenomena occuring jointly in these texts are currenlty tackled as separated tasks in V&L or text processing. We argue that a promising direction for future research in V&L is to aim for a joint framework that combines these different phenomena and levels of analysis. We believe that such a framework would be useful in a range of typical NLP applications (such as information extraction) where, currently, state-of-the-art models usually only process the text of a multimodal document. Arnold and Tilton (2020) discuss the motivation for such projects in the context Digital Humanities.

The example documents discussed in this paper differ from typical objects of V&L research in many respects, but most importantly in terms of their (i) structure and (ii) semantics or topic. Thus, our building articles are relatively long (i.e. much longer than image paragraphs in Krause et al. (2017)), contain multiple images and text segments that *do not* directly relate to any of the images. Concerning their semantics, the documents discuss buildings which constitute a type of named entity. This entity can be depicted visually in very diverse ways (see Section 2) and that can be associated with a rich body of knowledge (e.g. historical events) described in the text. We will show how these two aspects call for a V&L framework that accounts for *diverse referential relations* whereas in most V&L tasks assume a single, *fixed* text-image relation.

## 2 Qualitative Case Study

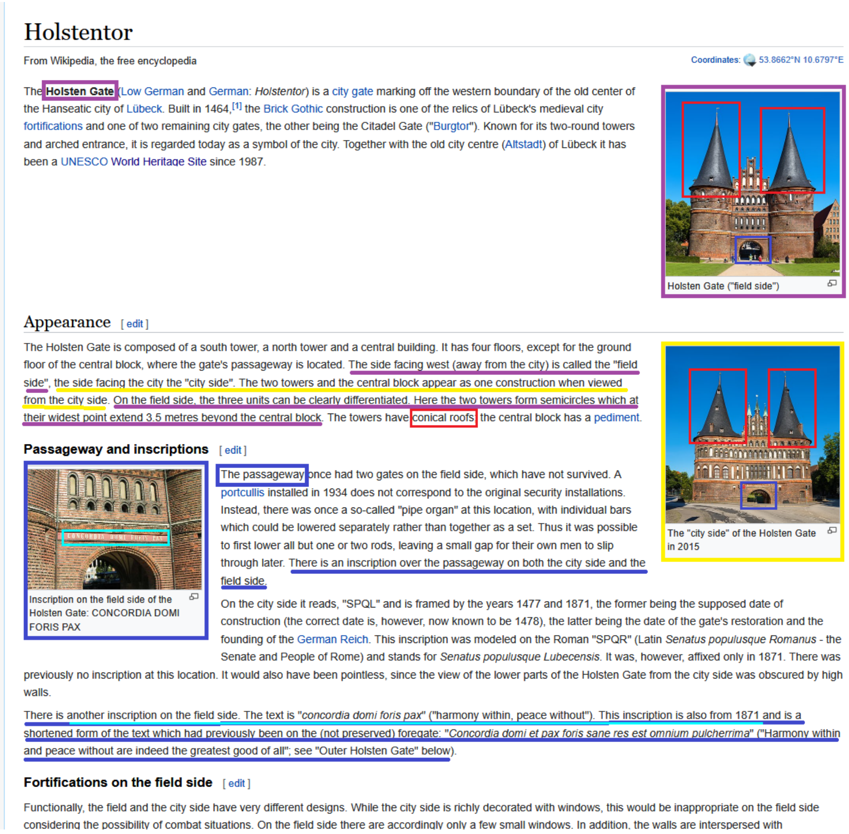
This section discusses observations we made by manually exploring a range of building articles from Wikipedia. We leave the empirical consolidation of our findings for future work.

### 2.1 Structure of Referential Relations

We first look at different structural aspects of referential relations in the *Holstentor* article, shown in Fig. 1. The first paragraph contains a general description of the entity, its location, importance etc., and is accompanied by a captioned image on its right side. In parallel to the text, this first image visually introduces the entity of the *Holstentor* but, other than that, has no further relations to paragraph it is aligned with.



*Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 53–60
Kyiv, Ukraine (Online), April 20, 2021



Holstentor

From Wikipedia, the free encyclopedia

Coordinates: 53.8662°N 10.6797°E

The **Holsten Gate** (Low German and German: *Holstentor*) is a city gate marking off the western boundary of the old center of the Hanseatic city of Lübeck. Built in 1464,[1] the Brick Gothic construction is one of the relics of Lübeck's medieval city fortifications and one of two remaining city gates, the other being the Citadel Gate ("Burgtor"). Known for its two-round towers and arched entrance, it is regarded today as a symbol of the city. Together with the old city centre (Altstadt) of Lübeck it has been a UNESCO World Heritage Site since 1987.

Holsten Gate ("field side")

Appearance [ edit ]

The Holsten Gate is composed of a south tower, a north tower and a central building. It has four floors, except for the ground floor of the central block, where the gate's passageway is located. The side facing west (away from the city) is called the "field side", the side facing the city the "city side". The two towers and the central block appear as one construction when viewed from the city side. On the field side, the three units can be clearly differentiated. Here the two towers form semicircles which at their widest point extend 3.5 metres beyond the central block. The towers have conical roofs, the central block has a pediment.

The "city side" of the Holsten Gate in 2015

**Passageway and inscriptions** [ edit ]

Inscription on the field side of the Holsten Gate: CONCORDIA DOMI FORIS PAX

The passageway once had two gates on the field side, which have not survived. A portcullis installed in 1934 does not correspond to the original security installations. Instead, there was once a so-called "pipe organ" at this location, with individual bars which could be lowered separately rather than together as a set. Thus it was possible to first lower all but one or two rods, leaving a small gap for their own men to slip through later. There is an inscription over the passageway on both the city side and the field side.

On the city side it reads, "SPQL" and is framed by the years 1477 and 1871, the former being the supposed date of construction (the correct date is, however, now known to be 1478), the latter being the date of the gate's restoration and the founding of the German Reich. This inscription was modeled on the Roman "SPQR" (Latin *Senatus populusque Romanus* - the Senate and People of Rome) and stands for *Senatus populusque Lubecensis*. It was, however, affixed only in 1871. There was previously no inscription at this location. It would also have been pointless, since the view of the lower parts of the Holsten Gate from the city side was obscured by high walls.

There is another inscription on the field side. The text is "*concordia domi foris pax*" ("harmony within, peace without"). This inscription is also from 1871 and is a shortened form of the text which had previously been on the (not preserved) foregate: "*Concordia domi et pax foris sane res est omnium pulcherrima*" ("Harmony within and peace without are indeed the greatest good of all"; see "Outer Holsten Gate" below).

**Fortifications on the field side** [ edit ]

Functionally, the field and the city side have very different designs. While the city side is richly decorated with windows, this would be inappropriate on the field side considering the possibility of combat situations. On the field side there are accordingly only a few small windows. In addition, the walls are interspersed with

Figure 1: First few paragraphs of the *Holstentor* article. Colour highlighting added for illustrating (approximate) text-image correspondences. Best viewed in colour.

This is completely different in the following paragraph which provides a detailed description of the building's appearance and is accompanied by a second image. The second image is visually very similar to the first: it shows the building in its entirety, but from a different perspective. The two opposing perspectives are explicitly referenced and explained in the *appearance* section, which even uses the same phrasing as the image captions. This paragraph also mentions parts of the gate, such as the *conical roofs*, which are shown in both images.

The first subsection of the *appearance* paragraph contains a third image that is spatially aligned with it. This subsection talks about the passageway; the image shows part (of one side) of it. Note that the caption refers to the inscription, which is located in the center of the aforementioned image. This inscription is first mentioned in the paragraph that is aligned with the image. Furthermore, the entire first paragraph *passageway* lists features of the building that are no longer visible in the contemporary photographs. The second paragraph talks about the other side of the gate, which is not pictured. The most relevant reference to the image of the passageway is contained in the final paragraph which describes its inscription in detail.

In sum, this example document shows (among other things) that discourse fragments of very different size (paragraph, sentences, sentence parts, noun phrases) can refer to images as well as their regions, in a way that can be difficult to disentangle.

### 2.2 Semantic Types of Referential Relations

While in the previous Section 2.1, we showed different cases for *what* text fragments can relate to an image, we now discuss examples for *how* these images relate to the text. In our qualitative case study on buildings, we observed 4 frequent relations - *generic* (view of the building), *related entity, detail/part and event.*, discussed below. This classification of relation types is not empirically validated and probably far from complete, but we intend to show that there are semantically very distinct types, even in the highly restricted building domain.

**Generic** Generic images show the general appearance of the building and lack a concrete reference to a specific part of the text, as, for instance, the first image in Fig. 1. These images appear at various points in the document and often contain the year the image was created in their caption. In most cases, they are arranged in chronological fashion, as in Fig. 2a[1].

**Related Entity** Images of related entities show people (or rarely, other buildings) that are relevant to the history of the building under discussion in the article. The entity depicted in the picture is almost always explicitly referenced in the article. For example, Fig. 2b has a named individual that is named in both the caption and article.

**Part or Detail** Images of the *parts* type show parts or details of the entity in question. This is the most diverse category in terms of the content of the images themselves. What is depicted can range from small details like plaques to major parts of the buildings like a tower, see Fig. 1. In some cases, these parts are not physically part of the building itself, but instead something that is permanently at the same location (e.g. an *organ* in a church).

**Event** Images in our building domain can also depict an *event* that takes place at the building in question. In the example in Fig. 2c, the image portrays the event in progress, e.g. a plane flying through the *Arc de Triomphe*. There is also another, more subtle but also frequent subtype of event-related image which relates to the existence of the architectural object itself and is often linked to its (partial or complete) destruction, construction or its renovation. In these cases, images often show the aftermath of the destruction (as is the case in Fig. 2d) or the site in the process of renovation/rebuilding. This latter case is particularly challenging in terms of semantic analysis and image-text matching, because it often entails scenarios where the text refers to (parts of) buildings that are no longer present in the image. That means, a model that fully and correctly analyses this scenario would need to make the connection between the text passages talking about the building's destruction and the image of it in a ruined state. Both types of events generally have a clear reference to the article's text, though the length can vary.

**Medium** In addition to the image content, we observe that the original medium is highly relevant in figuring out its semantic relation. The domain of buildings is especially rich in different original media - articles contain digitized images of paintings, sketches, maps, diagrams, post cards or photographs. To a human reader, these are not only understood differently, but themselves contain information.

### 2.3 Discussion

Inutitively, the Wikipedia articles on buildings that we have discussed in this section constitute a relatively simple type of multi-modal document: each document introduces and discusses a single, depictable, main entity, both in terms of textual and visual elements. Many images have captions that refer to the image's main object. Typically, this object is also referred to explicitly (often with very similar verbiage) in the text, except the tricky case of *generic images* where it is less clear which specific discourse fragment is referentially related.[2] Moreover, the articles have a clear hierarchical structure and, most of the time, images are positioned next to the paragraph that they are related to.

Formally, however, these examples indicate that there is a lot of structural and semantic complexity found in naturally occuring text-image correspondences. A long range of questions could be asked to capture this correspondence: (i) which text spans refer to an image or image region and which do not refer? (ii) what is the size of the text span that refers to the image (i.e. paragraph, sentence, noun phrase, ...), (iii) which text spans refer to the same image?, (iv) which images refer to the same text span or entity?, (v) how does the text refer to the image?, (vi) how do caption and text relate to each other?, etc. As we will discuss below, most existing V&L task come nowhere near this complexity and, most notably, make a lot of simplications on the text analysis side.

## 3 Related Work in Vision & Language

**Fixed Text-Image Relation** Probably the most widely used image-text data in Vision & Language are so-called image captions that provide a general, neutral description of an image's content, cf. MSCOCO (Lin et al., 2014) or Flickr30k (Plummer et al., 2015) captions . Typical captions consist of a single sentence that directly refers to the image. Other work has looked at more fine-grained referential relations such as referring expressions in the

[1]for Fig. 2, see supplementary material

[2]They could be seen as being linked to the text as a whole, but this is not particularly informative for information extraction or similar tasks.

form of noun phrases that identify specific objects in an image (Kazemzadeh et al., 2014). Work on even more fine-grained resolution that captures object parts is relatively rare, but see (Hürlimann and Bos, 2016). A complementary trend is to use texts that are (slightly) longer than image captions, such as image paragraphs that describe the image content in a sequence of sentences (Krause et al., 2017) or dialogues that center on identifying an object in a sequence of turns (de Vries et al., 2017) or an image from a set of images (Das et al., 2017). All of these datasets are crowdsourced and target a referential task on a specific, fixed level, i.e. image-sentence, object-phrase, object-dialogue, image-dialogue. It is worth noting that, internally, many recent large-scale models in V&L process object-phrase relations while encoding image-sentence pairs (Lee et al., 2018; Anderson et al., 2018; Kottur et al., 2018; Tan and Bansal, 2019; Lu et al., 2020), combining referential relations on two different levels. None of these tasks and models, however, deal with image-text pairs where significant parts of the text have no relation to the visual content, thereby circumventing the need to identify fragments that do indeed stand in a referential relation to a given image.

**Diverse Referential Relations** There is some initial work on datasets and tasks that capture more varied semantic or discursive relations between image and text: Kruk et al. (2019) tag the image intent in multi-modal Twitter posts, distinguishing between intents like 'provocative', 'expressive' or 'promotive'. Their annotations assign a global label to the image which captures the relation to the text as a whole. This goes beyond literal image descriptions, but still does not capture structurally diverse referential relations. Alikhani et al. (2019) investigate text-image coherence in recipe texts that describe sequences of consecutive actions in a cooking context. Structurally, the recipe's text is already segmented, with an image aligned to each step. Alikhani et al. (2019) distinguish image-text relations with respect to which part of the action is shown and whether all entities affected by an action are visible/mentioned in the text. Both papers work on naturally occuring text, though these are still relatively short (tweets and 1-2 sentences per step respectively). Neither task faces the segmentation problem to a degree that is similar to the complex structure we encountered in multi-modal Wikipedia articles. By contrast, in our building example, the rhetorical purpose and authorial intent of each picture seems to be more or less uniform. That is, images are included to illustrate (as opposed to being provocative or expressive). Likewise, the semiotics of these images are overwhelmingly parallel to the content of the text.

Muraoka et al. (2020) work with a more coarse-grained and somewhat simplified version of the problem discussed in this paper. Their task is to correctly predict the physical alignment of images and sections in Wikipedia articles. This approach utilizes the inherent document structure[3], however our observations (see Section 2) call into question the presupposition that alignment in layout entails alignment in content. A similar text-image matching task is discussed in Hessel et al. (2019), where the authors seek to match the images in a document to the most relevant sentences in it (leaving out the captions). Their model is trained on collections of sentences and images from the same documents; or different documents, for instances of non-relatedness. This information is used at test time to estimate the individual links between the sentences and images of a given document. Hessel et al. (2019) is highly relevant to the concerns discussed in this paper because it has some success in grappling with the comparatively large amounts of text in the Wikipedia article genre.

## 4 Towards Multi-Modal Information Extraction

Many of the phenomena we have discussed in Section 2 have long been researched in NLP models that extract information from text only. Prominently, text-oriented NLP has long been interested in detecting and processing entities, i.e. Named Entity Recognition (NER) is a very well-known NLP task that is useful in a range of applications (Li et al., 2020). The most standard named entity categories are `person`, `location` and `organisation`, however there is a number of NER tools with varying tag sets. One way to model texts of the type illustrated in Fig. 1 would be to move towards multi-modal NER models that identify mentions of entities in a text and link them to corresponding images or image regions, cf. Asgari-Chenaghlu et al. (2020) for a similar proposal.

Event detection is another text-based NLP task that has been approached with the use of CNNs (Nguyen and Grishman, 2015) and, more recently,

[3] and consequently save on expensive manual annotation

attention mechanisms (Liu et al., 2017b). Multi-modal event detection could be useful to capture referential relations as shown in Fig. 2c. Finally, models that represent or encode relations between entities (Lin et al., 2016; Zhang et al., 2019) in a multi-modal text would be an extremely useful tool in our setting. As a step towards processing comparatively large chunks of text, *discourse segmentation* (Braud et al. (2017), Iruskieta et al. (2019)) splits documents into elementary Discourse Units. Parsing these texts as a discourse is also a topic of ongoing research (Liu et al., 2017a; Li et al., 2016).

To the best of our knowledge, research in Vision & Language has hardly been inspired by these classical, entity-centric task in text processing. This general impression is corroborated by the very comprehensive V&L survey of Mogadala et al. (2019).

## 5 Conclusion

In this paper, we have discussed some complexities of referential relations that arise in natually occurring multi-modal texts. Solving at least some of these requires the use of far more involved text processing techniques than is common for widespread V&L tasks such as image captioning or visual dialogue. Our domain - architectural sites - narrows this potentially sprawling problem somewhat. While every building is an entity unto itself, there are common features that are shared by large subsets. We argue that Wikipedia articles are a valuable source of raw data for multi-modal document analysis, since they constitute a genre of document that is freely available in large quantities and across languages.[4] However, it is questionable whether models that identify the type of image-text relations discussed in this paper can be developed without hand-annotated data. In terms of uitlity and intended audience, it may be worth considering work like Arnold and Tilton (2020), whose aim is to add robust, searchable annotations to existing collections of historical images. This leads the authors to develop a model that automatically labels images using image segmentation and a pre-defined ontology. We believe that moving towards such more realistic texts in V&L is interesting both from a linguistic, and from an application-oriented perspective, i.e. for multi-modal information extraction.

[4] Some tasks and datasets may also benefit from existing knowledge bases such as Wikidata (Vrandečić and Krötzsch, 2014).

## Acknowledgements

This work was supported by a grant from the Federal Ministry of Education and Research (BMBF, grant No. 01UG2120A).

## References

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 570–575, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Taylor Arnold and Lauren Tilton. 2020. Enriching historic photography with structured data using image region segmentation. In *Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access*, pages 1–10, Marseille, France. European Language Resources Association (ELRA).

Meysam Asgari-Chenaghlu, M. Reza Feizi-Derakhshi, Leili Farzinvash, M. A. Balafar, and Cina Motamed. 2020. A multimodal deep learning approach for named entity recognition from social media.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 237–243, Vancouver, Canada. Association for Computational Linguistics.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960.

Judy Delin and John Bateman. 2002. Describing and critiquing multimodal documents. *Document Design*, 3:140–155.

Michel Hardy-Vallée. 2016. Text and image: a critical introduction to the visual/verbal divide by john a. bateman. *Visual Studies*, 31:366–368.

Jack Hessel, Lillian Lee, and David Mimno. 2019. Unsupervised discovery of multimodal links in multi-image, multi-sentence documents. 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)*, pages 2034–2045, Hong Kong, China. Association for Computational Linguistics.

Manuela Hürlimann and Johan Bos. 2016. Combining lexical and spatial knowledge to predict spatial relations between objects in images. In *Proceedings of the 5th Workshop on Vision and Language*, pages 10–18.

Mikel Iruskieta, Kepa Bengoetxea, Aitziber Atutxa Salazar, and Arantza Diaz de Ilarraza. 2019. Multilingual segmentation based on neural networks and pre-trained word embeddings. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 125–132, Minneapolis, MN. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. 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)*, pages 4622–4632, Hong Kong, China. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

J. Li, A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017a. Learning character-level compositionality with visual features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2068, Vancouver, Canada. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017b. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Vancouver, Canada. Association for Computational Linguistics.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2019. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.

Masayasu Muraoka, Ryosuke Kohita, and Etsuko Ishii. 2020. Image position prediction in multimodal documents. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4265–4274, Marseille, France. European Language Resources Association.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371.

Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. 2020. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. *arXiv preprint arXiv:2007.02375*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

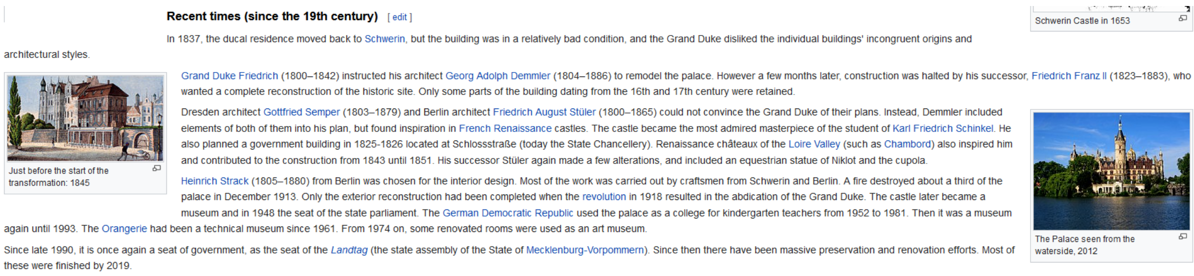
Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40. Language in Vision.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5356–5362.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

# 6 Supplemental Material



Recent times (since the 19th century) [ edit ]

Schwerin Castle in 1653

In 1837, the ducal residence moved back to Schwerin, but the building was in a relatively bad condition, and the Grand Duke disliked the individual buildings' incongruent origins and architectural styles.

Just before the start of the transformation: 1845

Grand Duke Friedrich (1800–1842) instructed his architect Georg Adolph Demmler (1804–1886) to remodel the palace. However a few months later, construction was halted by his successor, Friedrich Franz II (1823–1883), who wanted a complete reconstruction of the historic site. Only some parts of the building dating from the 16th and 17th century were retained.

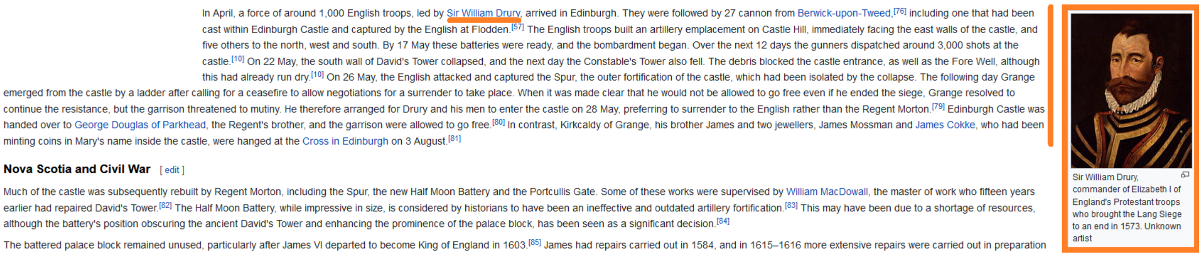
Dresden architect Gottfried Semper (1803–1879) and Berlin architect Friedrich August Stüler (1800–1865) could not convince the Grand Duke of their plans. Instead, Demmler included elements of both of them into his plan, but found inspiration in French Renaissance castles. The castle became the most admired masterpiece of the student of Karl Friedrich Schinkel. He also planned a government building in 1825-1826 located at Schlossstraße (today the State Chancellery). Renaissance châteaux of the Loire Valley (such as Chambord) also inspired him and contributed to the construction from 1843 until 1851. His successor Stüler again made a few alterations, and included an equestrian statue of Niklot and the cupola.

Heinrich Strack (1805–1880) from Berlin was chosen for the interior design. Most of the work was carried out by craftsmen from Schwerin and Berlin. A fire destroyed about a third of the palace in December 1913. Only the exterior reconstruction had been completed when the revolution in 1918 resulted in the abdication of the Grand Duke. The castle later became a museum and in 1948 the seat of the state parliament. The German Democratic Republic used the palace as a college for kindergarten teachers from 1952 to 1981. Then it was a museum again until 1993. The Orangerie had been a technical museum since 1961. From 1974 on, some renovated rooms were used as an art museum.

The Palace seen from the waterside, 2012

Since late 1990, it is once again a seat of government, as the seat of the *Landtag* (the state assembly of the State of Mecklenburg-Vorpommern). Since then there have been massive preservation and renovation efforts. Most of these were finished by 2019.

(a) Generic: Two chronologically arranged generic images in the *Schwerin Castle* article



In April, a force of around 1,000 English troops, led by Sir William Drury, arrived in Edinburgh. They were followed by 27 cannon from Berwick-upon-Tweed,[76] including one that had been cast within Edinburgh Castle and captured by the English at Flodden.[57] The English troops built an artillery emplacement on Castle Hill, immediately facing the east walls of the castle, and five others to the north, west and south. By 17 May these batteries were ready, and the bombardment began. Over the next 12 days the gunners dispatched around 3,000 shots at the castle.[10] On 22 May, the south wall of David's Tower collapsed, and the next day the Constable's Tower also fell. The debris blocked the castle entrance, as well as the Fore Well, although this had already run dry.[10] On 26 May, the English attacked and captured the Spur, the outer fortification of the castle, which had been isolated by the collapse. The following day Grange emerged from the castle by a ladder after calling for a ceasefire to allow negotiations for a surrender to take place. When it was made clear that he would not be allowed to go free even if he ended the siege, Grange resolved to continue the resistance, but the garrison threatened to mutiny. He therefore arranged for Drury and his men to enter the castle on 28 May, preferring to surrender to the English rather than the Regent Morton.[79] Edinburgh Castle was handed over to George Douglas of Parkhead, the Regent's brother, and the garrison were allowed to go free.[80] In contrast, Kirkcaldy of Grange, his brother James and two jewellers, James Mossman and James Cokke, who had been minting coins in Mary's name inside the castle, were hanged at the Cross in Edinburgh on 3 August.[81]

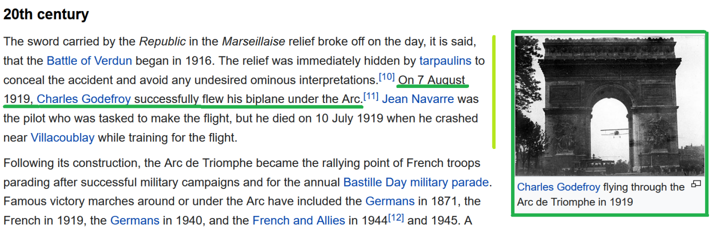
Sir William Drury, commander of Elizabeth I of England's Protestant troops who brought the Lang Siege to an end in 1573. Unknown artist

Nova Scotia and Civil War [ edit ]

Much of the castle was subsequently rebuilt by Regent Morton, including the Spur, the new Half Moon Battery and the Portcullis Gate. Some of these works were supervised by William MacDowall, the master of work who fifteen years earlier had repaired David's Tower.[82] The Half Moon Battery, while impressive in size, is considered by historians to have been an ineffective and outdated artillery fortification.[83] This may have been due to a shortage of resources, although the battery's position obscuring the ancient David's Tower and enhancing the prominence of the palace block, has been seen as a significant decision.[84]

The battered palace block remained unused, particularly after James VI departed to become King of England in 1603.[85] James had repairs carried out in 1584, and in 1615–1616 more extensive repairs were carried out in preparation

(b) Related Entity: Portrait of an aristocrat in the article of *Edinburgh Castle*



20th century

The sword carried by the *Republic* in the *Marseillaise* relief broke off on the day, it is said, that the Battle of Verdun began in 1916. The relief was immediately hidden by tarpaulins to conceal the accident and avoid any undesired ominous interpretations.[10] On 7 August 1919, Charles Godefroy successfully flew his biplane under the Arc.[11] Jean Navarre was the pilot who was tasked to make the flight, but he died on 10 July 1919 when he crashed near Villacoublay while training for the flight.

Charles Godefroy flying through the Arc de Triomphe in 1919

Following its construction, the Arc de Triomphe became the rallying point of French troops parading after successful military campaigns and for the annual Bastille Day military parade. Famous victory marches around or under the Arc have included the Germans in 1871, the French in 1919, the Germans in 1940, and the French and Allies in 1944[12] and 1945. A

(c) Event: Flight through the *Arc de Triomphe*

Destruction and rebuilding [ edit ]

The building was mostly destroyed by the carpet bombing raids of 13–15 February 1945. The art collection had been previously evacuated, however. Reconstruction, supported by the Soviet military administration, began in 1945; parts of the restored complex were opened to the public in 1951. By 1963 the Zwinger had largely been restored to its pre-war state.

Ruins of the wall pavilion in 1945.

See also [ edit ]

- Pillnitz Castle – Summer residence of the electors and kings of Saxony
- Moritzburg Castle – Hunting lodge of the electors and kings of Saxony
- List of castles in Saxony
- List of Baroque residences

References [ edit ]

1. ^ Gurlitt: *Kunstdenkmäler Dresdens*, H. 2, p. 313

(d) Event: Destruction of the *Zwinger, Dresden*

Figure 2: Examples of Semantic Referential Relations