SzegedKoref: A Hungarian Coreference Corpus

Veronika Vincze[1,2], **Klára Hegedűs**[3], **Alex Sliz-Nagy**[4], **Richárd Farkas**[4]
[1]MTA-SZTE Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of General Linguistics, University of Szeged
[3]Department of Psychology, University of Szeged
klarahegedus92@gmail.com
[4]Institute of Informatics, University of Szeged
rfarkas@inf.u-szeged.hu
a.sliz.nagy@gmail.com

Abstract

In this paper we introduce SzegedKoref, a Hungarian corpus in which coreference relations are manually annotated. For annotation, we selected some texts of Szeged Treebank, the biggest treebank of Hungarian with manual annotation at several linguistic layers. The corpus contains approximately 55,000 tokens and 4000 sentences. Due to its size, the corpus can be exploited in training and testing machine learning based coreference resolution systems, which we would like to implement in the near future. We present the annotated texts, we describe the annotated categories of anaphoric relations, we report on the annotation process and we offer several examples of each annotated category. Two linguistic phenomena – phonologically empty pronouns and pronouns referring to subordinate clauses – are important characteristics of Hungarian coreference relations. In our paper, we also discuss both of them.

Keywords: coreference, corpus, Hungarian

1. Introduction

In order to avoid unnecessary repetitions and redundancy, speakers can use a wide variety of expressions when referring to the same entity or event in the world. Languages usually offer several lexical and grammatical tools for this purpose. One of the grammatical tools to express identity is coreference, which is used when two (or more) linguistic units refer to the same entity/individual in the world. Coreference relations are most frequently expressed by pronouns, adverbs and nouns (mostly, nouns denoting gender or position such as *girl* or *sergeant*). At the lexical level, it is mostly synonyms that can contribute to lexical variability.

In this paper we introduce the SzegedKoref corpus, in which coreference relations are manually annotated. For annotation, we selected some texts of Szeged Treebank. It is the biggest treebank of Hungarian that contains manual annotation at several linguistic layers (Csendes et al., 2005). We present the annotated texts, we describe the annotated categories of anaphoric relations – pronominal, nominal, adverbial and verbal coreference and subtypes of nominal coreference (hypernyms, synonyms etc.), and we offer several examples of each annotated category. We also mark zero anaphors and pronouns coreferential with subordinate clauses since these are two linguistic phenomena of Hungarian that deserve special attention from the viewpoint of coreference resolution.

2. Related Work

There are several coreference corpora available for many languages, for instance, OntoNotes contains coreference annotation for English, Chinese and Arabic (Weischedel et al., 2011; Pradhan et al., 2007). This database formed the training and test sets of the CoNLL-2011 (Pradhan et al., 2011) and CoNLL-2012 (Pradhan et al., 2012) shared tasks, which aimed at automatic coreference resolution.

There is coreference annotation in the DIRNDL and ANCOR_Centre corpora, containing German and French spoken language data (Muzerelle et al., 2014; Björkelund et al., 2014). As for Japanese, the corpus NAIST Text contains coreference annotation, together with predicate-argument structure (Iida et al., 2007). A large coreference corpus is also available for Polish (Ogrodniczuk et al., 2014; Ogrodniczuk et al., 2013b), moreover, there are annotated coreference corpora for Dutch (Hendrickx et al., 2008) and Czech (Nedoluzhko et al., 2009) as well. Recently, Ghaddar and Langlais (2016) reported on WikiCoref, a coreference corpus of English Wikipedia articles.

A small dataset with manual coreference annotation was earlier published for Hungarian (Miháltz, 2012). In contrast, here we present our large corpus, SzegedKoref, which has been manually annotated for coreference data. Due to its size, the corpus can be used for training and evaluating machine learning-based systems, which is nowadays the most popular approach used for coreference resolution (Pradhan et al., 2012).

In morphologically rich languages like Hungarian, some issues might occur concerning the annotation process of coreference relations. It is the treatment of phonologically empty pronouns that is particularly important among others, as already emphasized for Polish (Ogrodniczuk et al., 2013a). Moreover, pronouns referring to subordinate clauses should also be paid special attention in Hungarian. In our paper, we will focus on both of these phenomena.

3. The Corpus

As the Szeged Corpus (Csendes et al., 2005) contains annotation for several linguistic layers (POS-tags, constituency and dependency syntax), we selected those texts for coreference annotation, in order to enrich their linguistic struc-

ture. Since it is preferred to annotate coreference relations in longer comprehensive texts instead of using very short texts, we also needed to select the appropriate subcorpora of the Szeged Corpus. For this reason, we finally decided to neglect the subcorpus containing short business news, where each piece of news consisted of only 1-2 sentences, hence annotation was not carried out in this subcorpus. Instead, we chose to focus on student essays and newspaper articles, which are comprehensive texts of considerable length and are expected to contain various coreference relations.

3.1. Annotation Principles

During annotation, mentions (i.e. mostly noun phrases that refer to a concept) were first marked, then antecedents were linked to the heads referring to the same entity. The type of coreference is also marked in the data, that is, pronominal, nominal, adverbial and verbal coreference. We also paid attention to derivational anaphors, i.e. cases where the antecedent and the head refer to the same action/entity but belong to different parts of speech (for instance, an action is expressed by a verb first, then it is referred to with a noun or participle). Categories are shown below:

- pronominal anaphor: *Ismertem **a lányt**, **aki** épp átjött az úton.* "I knew **the girl who** was just crossing the street."
- nominal anaphor:
  - Repetition: *Józsi este találkozott **a lánnyal**. **A lány** piros ruhát viselt.* "Joe met **the girl** last night. **The girl** was wearing a red dress."
  - Variant: ***Pálffy János gróf** személyében magyar főparancsnokot neveztek ki a császári sereg élére. **Pálffy** tárgyalásokat kezdett Károlyi Sándor báróval.* "A Hungarian colonel – **Earl János Pálffy** – was chosen to lead the imperial army. **Pálffy** initiated negotiations with Baron Sándor Károlyi."
  - Synonym: *Józsi kapott **egy biciklit**. Másnap **az új kerékpárral** jött munkába.* "Joe got **a new bike**. The next day he came to work with **his new bicycle**."
  - Hypernym: *Az udvaron volt **egy kutya**. **Az állat** keservesen ugatott.* "There was **a dog** in the yard. **The animal** was barking desperately."
  - Hyponym: *Az udvaron volt **egy kutya**. **Szegény uszkár** meg volt kötve.* "There was **a dog** in the yard. **The poor poodle** was tied."
  - Meronym: *Jól játszott **a csapat**, **a kapus** különösen kiemelkedett a mezőnyből.* "**The team** was playing well, **the goalkeeper** especially had an excellent performance."
  - Holonym: *Defektes lett **a jobb első kerék**, így **az autónak** ki kellett állnia a versenyből.* "**The first right wheel** got a puncture, so **the car** had to finish the race."
  - Epithet: ***Józsi** nem tudott bejutni, mert **a szerencsétlen** otthon hagyta a kulcsot.* "**Joe** could not enter the flat because **the poor one** forgot his key at home."
  - Apposition: *Pálffy tárgyalásokat kezdett **Rákóczi megbízottjával, Károlyi Sándor báróval**.* "Pálffy initiated negotiations with **Rákóczi's representative**, **Baron Sándor Károlyi**."
- adverbial anaphor: *Elindultunk **a hotelba**, a többiekkel **ott** találkozunk.* "We have left for **the hotel**, we will meet the others **there**."
- verbal anaphor: *Juli **elénekelt** tegnap egy dalt, ma pedig Józsi is **így tett**.* "Julie **sang a song** yesterday, and Joe **did so** today."
- derivational anaphor: *Józsi mindig **énekel** a fürdőben. **Az éneklés** nagyon zavarja a többi lakót.* "Joe always **sings** in the bathroom. **His singing** annoys the other tenants."

As for nominal anaphors, we also marked their semantic categories, for instance, whether there is a synonymy/hypernymy/holonymy relation between the head and mention (e.g. *kutya* "dog" – *állat* "animal"), whether the head is simply repeated (e.g. *kutya* "dog" – *kutya* "dog") or whether a variant is used (e.g. *Albert Einstein – Einstein*). Derivational relations were also marked between the head and the anaphor (e.g. *Pista hangosan **énekelt**. Az **ének** nagyon zavarta a szomszédját.* "Steve was **singing** loudly. His **song** annoyed his neighbour.").

In Hungarian, zero pronouns also mean a challenge to coreference resolution systems. As the Hungarian verbal paradigm differentiates between verb forms referring to a definite object and verb forms referring to an indefinite one on the one hand, and verbs are also conjugated differently for each person and number on the other hand, there is no need to explicitly mark pronominal subjects and objects in the sentence and so, they can be deduced from context. Furthermore, pronominal possessors might also remain hidden in possessive constructions, due to nominal inflection. From the viewpoint of coreference resolution, all this entails that the anaphor might not be present in the sentence as a separate token, only as a zero pronoun (pro). Thus, before the annotation process started, they had had to be inserted into the text. The following example illustrates this process:

> *Látta a kertjében.* → ***proSUBJ** látta **proOBJ** a **proPOSS** kertjében.*
> see-PAST-3SGOBJ the garden-3SGPOSS-INE
> → proSUBJ see-PAST-3SGOBJ proOBJ the proPOSS garden-3SGPOSS-INE
> "He saw it in his garden."

Here, the words equivalent to the English pronouns *he*, *it* and *his* are missing from the original Hungarian sentence and instead, zero pronouns were automatically inserted into the text before the manual annotation process, so they are also annotated in the data. The insertion took place on the

basis of linguistic rules and morphological and syntactic constraints.
Pronouns referring to subordinate clauses were also marked as coreferent with the subordinate clause they are referring to, no matter they occurred in their overt or zero form. In contrast with English, Hungarian may use a pronoun in the matrix clause that can function as an argument of the main verb and is coreferent with the subordinate clause. For instance, compare these two sentences:

> *Mondtam **proOBJ**, hogy mindjárt itt a karácsony.*
> say-PAST-1SGOBJ proOBJ , that soon here the Christmas
> "I told you that Christmas is almost here."

> ***Azt** mondtam, hogy mindjárt itt a karácsony.*
> it-ACC say-PAST-1SGOBJ , that soon here the Christmas
> "I told you that Christmas is almost here." (lit. "I told **it** to you that Christmas is almost here."

In these sentences, the overt pronoun *azt* and the zero pronoun *proOBJ* were annotated as coreferent with the clause *mindjárt itt a karácsony*.

3.2. Annotation Process

Annotation was carried out by two annotators, who were trained in linguistics and supervised by a linguist expert. The MMAX2 tool was employed for annotation, which allows multilayer annotation and makes it possible to visually track coreference chains during annotation (Müller and Strube, 2006). A sample of the annotated texts is shown in Figure 1.
In order to measure inter-annotator agreement rate, a small sample of 10 documents were annotated by both annotators. Their agreement rate was 0.95 (in terms of F-score), with regard to mention identification.

3.3. Statistical Data

Currently, the corpus contains 309 sentences and 9,782 tokens from the newspaper domain and 3,712 sentences and 45,981 tokens from the student essay subcorpus. Altogether, there are 400 texts, 4021 sentences and 55,763 tokens in the current version of the corpus.
There are 2191 anaphoric chains in the student essay subcorpus and 265 in the newspaper domain, adding up to 2456 anaphoric chains altogether. As shown in Table 1, the most frequent types of anaphor are pronominal anaphors and repetition, indicating that automatic coreference resolution systems should pay extra attention to these categories. Figure 2 tells us that repetitions, hypernyms and adverbial anaphors are much more frequent in the student essays than in the newspaper articles. However, synonyms and appositions are more widely applied in newspaper texts.
The distribution of the anaphoric categories shows a statistically significant difference (χ^2-test, $p<0.01$), hence there are domain differences in the use of anaphoric categories. Later on, we intend to annotate other domains of texts for coreference in order to check what the most characteristic anaphoric categories are for each domain.

Zero pronoun	**Student essays**	**Newspaper**	**Total**
subject	594	119	713
object	181	9	190
possessive	212	128	340
Total	987	256	1243

Table 2: Anaphoric zero pronouns.

Table 2 shows that there are many zero pronouns that form part of an anaphoric chain, what is more, about 67% of pronominal anaphors involve a zero pronoun. Hence, coreference resolution systems should be prepared for the efficient treatment of Hungarian zero pronouns.

4. Possible Uses of the Corpus

Coreference corpora and coreference resolution algorithms might be useful for several purposes. For instance, information extraction systems might exploit coreference relations, since information related to a specific entity might be collected from the text not only by searching for the exact name of the entity but also by finding elements that are coreferent with it.
On the other hand, machine translation applications might also profit from coreference resolution. Although Hungarian does not make use of a grammatical gender for nouns and pronouns, it can be essential to know whether a given pronoun (e.g. *őt* "him" or "her") refers to a male or female person as this information is crucial in finding the proper equivalent of the pronoun in another language that uses grammatical gender. With the antecedent of the pronoun identified, the system may be able to select the personal pronoun of the appropriate gender.
The thorough investigation of types of coreference, as well as the detailed analysis of zero pronouns and pronouns referring to clauses, might be also fruitful for both theoretical linguistics and natural language processing.

5. Conclusions

Here we introduced the SzegedKoref corpus, in which coreference relations are manually annotated. The corpus contains selected texts of Szeged Treebank, the biggest treebank of Hungarian with manual annotation at several linguistic layers. We presented the basic annotation principles and some statistical data on the annotated corpus. Due to its size, the corpus can be exploited in training and testing machine learning based coreference resolution systems, which we would like to implement in the near future.
The corpus is freely available for research and educational purposes at `http://rgai.inf.u-szeged.hu/SzegedTreebank`.

6. Acknowledgements

This research was supported by the project "Integrated program for training new generation of scientists in the fields of computer science", no EFOP-3.6.3-VEKOP-16-2017-0002 and further supported by the EU-funded Hungarian grant EFOP-3.6.1-16-2016-00008. The project has been supported by the European Union and co-funded by the European Social Fund.

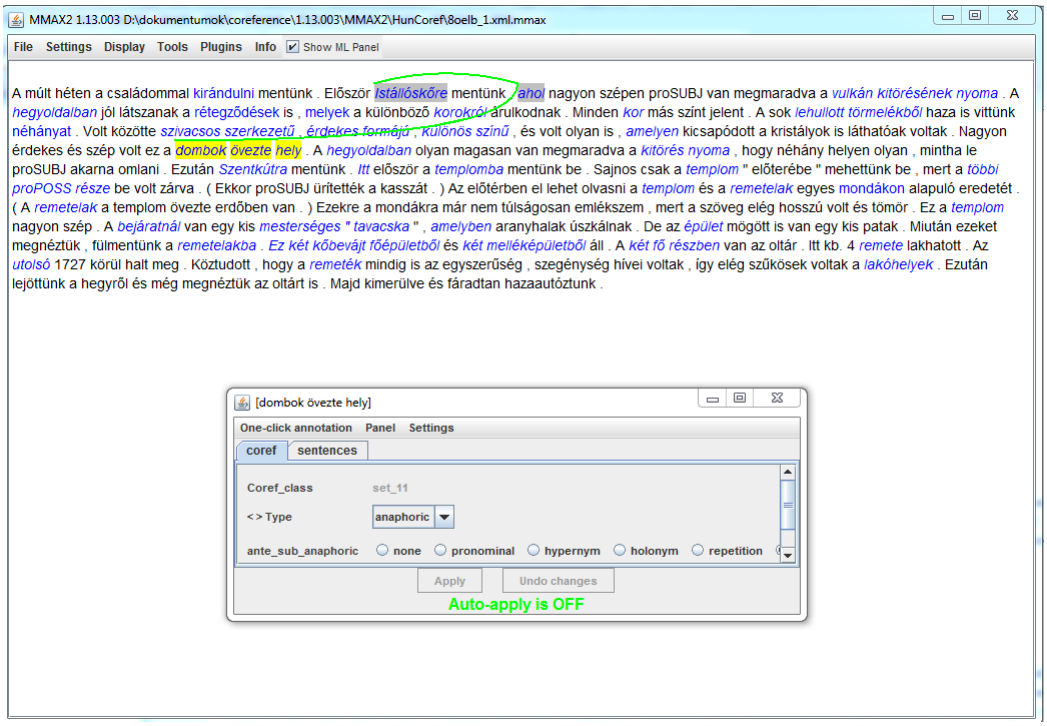

Figure 1: The MMAX2 annotation tool.

Anaphor	Student essays	%	Newspaper	%	Total	%
pronominal	1531	33.51	320	39.22	1851	34.37
repetition	1176	25.74	86	10.54	1262	23.44
synonym	329	7.20	252	30.88	581	10.79
hypernymy	445	9.74	0	0.00	445	8.26
holonymy	350	7.66	34	4.17	384	7.13
epitheton	17	0.37	23	2.82	40	0.74
apposition	117	2.56	70	8.58	187	3.47
adverbial	339	7.42	1	0.12	340	6.31
verbal	5	0.11	0	0.00	5	0.09
derivational	76	1.66	30	3.68	106	1.97
other	184	4.03	0	0.00	184	3.42
Total	4569	100	816	100	5385	100

Table 1: Types and frequency of anaphors.

The research of Richárd Farkas was funded by the János Bolyai Scholarship. Veronika Vincze was supported by the UNKP-17-4 New National Excellence Program of the Ministry of Human Capacities.

7. Bibliographical References

Björkelund, A., Eckart, K., Riester, A., Schauffler, N., and Schweitzer, K. (2014). The Extended DIRNDL Corpus as a Resource for Coreference and Bridging Resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3222–3228, Reykjavik, Iceland. European Language Resources Association (ELRA).

Csendes, D., Csirik, J., Gyimóthy, T., and Kocsor, A. (2005). The Szeged TreeBank. In Václav Matousek, et al., editors, *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005*, Lecture Notes in Computer Science, pages 123–132, Berlin / Heidelberg, September. Springer.

Ghaddar, A. and Langlais, P. (2016). Wikicoref: An english coreference-annotated corpus of wikipedia articles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Vloet, J. V. D., and Verschelde, J.-L. (2008). A Coreference Corpus and Resolution System for Dutch. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007).

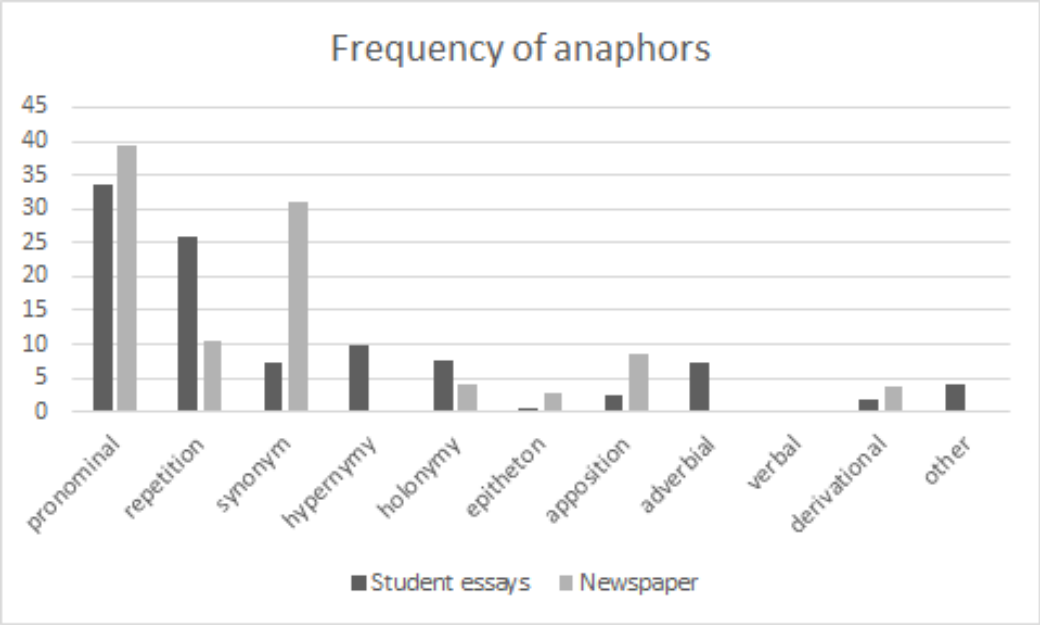

Figure 2: Frequency of anaphor types.

Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. In *Proceedings of the Linguistic Annotation Workshop*, pages 132–139, Prague, Czech Republic, June. ACL.

Miháltz, M. (2012). Tudásalapú koreferencia- és birtokosviszony-feloldás magyar szövegekben. *Általános Nyelvészeti Tanulmányok*, XXIV:151–166.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, et al., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., and Villaneau, J. (2014). ANCOR_Centre, a large free spoken French coreference corpus: description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 843–847, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nedoluzhko, A., Mírovskỳ, J., Ocelák, R., and Pergler, J. (2009). Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, pages 1–16.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2013a). Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In Maosong Sun, et al., editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, volume 8202 of *Lecture Notes in Computer Science*, pages 97–108. Springer Berlin Heidelberg.

Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., and Zawisławska, M. (2013b). Polish coreference corpus. In *6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, volume 3, pages 494–498. Wydawnictwo Poznańskie.

Ogrodniczuk, M., Kopeć, M., and Savary, A. (2014). Polish Coreference Corpus in Numbers. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3234–3238, Reykjavik, Iceland. European Language Resources Association (ELRA).

Pradhan, S. S., Ramshaw, L., Weischedel, R. M., MacBride, J., and Micciulla, L. (2007). Unrestricted Coreference: Identifying Entities and Events in OntoNotes. In *ICSC*, pages 446–453. IEEE Computer Society.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., and Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., and Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Weischedel, R., Hovy, E., Marcus, M., Palmer, M., Belvin, R., Pradan, S., Ramshaw, L., and Xue, N. (2011). OntoNotes: A Large Training Corpus for Enhanced Processing. In *Handbook of Natural Language Processing and Machine Translation*.