Leveraging Morpho-semantics for the Discovery of Relations in Chinese Wordnet

Shu-Kai Hsieh
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
shukaihsieh@ntu.edu.tw

Yu-Yun Chang
Graduate Institute of Linguistics
National Taiwan University
Taipei, Taiwan
yuyun.unita@gmail.com

Abstract

Semantic relations of different types have played an important role in wordnet, and have been widely recognized in various fields. In recent years, with the growing interests of constructing semantic network in support of intelligent systems, automatic semantic relation discovery has become an urgent task. This paper aims to extract semantic relations relying on the *in situ* morpho-semantic structure in Chinese which can dispense of an outside source such as corpus or web data. Manual evaluation of thousands of word pairs shows that most relations can be successful predicted. We believe that it can serve as a valuable starting point in complementing with other approaches, which will hold promise for the robust lexical relations acquisition.

1 Introduction

Semantic relations are at the core of WordNet-alike architecture, and constitute the essential and integral part of linguistic and conceptual knowledge formalization. However, the manual labeling task of semantic relations is very laborious.

To minimize the labor, in recent years, automatic ways of extracting semantic relations from textual data have been proposed. Among these methods, extensive works have been done based on the so-called *pattern-based* approaches, which was pioneered by (Hearst, 1992). The patterns predefined or plucked out of a corpus are often referred to as *lexico-syntactic patterns*, which serve as an information marker for a certain relation between two concepts. Later representative works using such approaches include (Cimiano et al., 2005), and (Pantel and Pennacchiotti, 2006), etc. Pattern-based extraction has shown quite reasonable success characterized by a (relatively) high precision rate, but suffers from a very low recall resulting from the fact that the patterns are rare in corpora. Remedies against the problem involve exploiting scaled data from the web (Cimiano et al., 2005), but runs the risk of influenced by the web genre (Alain, 2010).

To enrich the relations coverage in Chinese Wordnet (CWN), in this paper, we propose an *in situ* approach by exploiting the morph-semantic information. This method, simple and straightforward as it seems, does not incur the difficulties associated with *lexical gaps* in cross-language mapping that any translation-based model would encounter; and it is also economic and complementary with previous approaches in that we can dispense of an outside corpus resource.

In what follows, Section 2 gives a brief summary of lexical semantic relations acquisition from two perspectives. Section 3 explains the proposed methods for the automatic discovery of semantic relations, which are the main focus of this study. Section 4 shows the experiment results and discussion. Finally, we conclude this paper in Section 5.

2 Relations in Chinese Wordnet

Modelling on English WordNet, CWN has been launched by Academia Sinica in 2006 and continuously broadened its scope (Huang et al., 2010).[1]. The initial version of CWN contains a manually created fine-grained senses repository but sparse relations. However, semantic relation labeling is a time-consuming and labor-demanding task. Two main methods were employed to automatic relation acquisition.

2.1 Bilingual Bootstrapping Approach

Though lexical semantic relations (LSRs) could be presumed to be *more* universal than word senses in human languages, a direct

[1]Freely available at http://lope.linguistics.ntu.edu.tw/cwn

copying or simple porting of LSRs from one wordnet to another could possibly lead to invalid relations in the target wordnet. A broader view on the underlying inference logic of cross-language LSRs with 26 rules was first proposed by (Huang et al., 2002) and formally introduced in (Hsieh, 2009). A series of large-scaled bilingual bootstrapping experiments showed substantial improvements (with 55% precision) over baseline model (47%). However, it was also reported that among the correctly predicted LSRs, a large portion (c.a. 60%) belongs to *non-lexical relations* such as `similar to, pertainym, also see`, etc.

To look deeper into the issues, second experiment focusing only on the `hypernymy-troponymy` among the verbs was conducted. The bootstrapping model returned totally 12214 verb pairs mapped from WordNet 3.0, which were manually evaluated. The analysis shows that around 50% verb pairs can be recognized as fit in CWN, however, two main error types are identified: [1] Lexicalization of verbs: similar to the problems of lexical gap appeared in the cross-language sense mapping, a single word in English often has meanings that require several words in Chinese to explain. By analyzing the results, it is found that many verbs could not be described by a single lexeme in Chinese. [2] Mismatch of synset: other than the above, there are cases when the hypernymy-troponymy relations of the verb pairs are approved, but the synset that CWN chooses is not the same with that of PWN. This could be due to the different semantic ranges between CWN and PWN hypernymy-troponymy pairs, or due to the subtlety of sense division when the sense levels are similar.

The bilingual bootstrapping experiments showed that lexical relations turn out to be not subject to automatic importing and would still require tremendous human efforts of validations.

2.2 Pattern-based Approach

There has been a variety of studies on the automatic acquisition of lexical semantic relations, Hearst (Hearst, 1992) first proposed a *lexico-syntactic pattern* based method for automatic acquisition of hyponymy from unrestricted texts, and since then automatically finding semantic relations by using various pattern-based algorithm has become the most common approach.

We (Lo et al., 2008) have tried to define some patterns (e.g., *a manner of*) to extract troponymy among verbs in Chinese. To avoid the interference of unnecessary contextual information which may include modal verbs, hedging, negation that often occur in different corpus genres, we applied the proposed patterns on the gloss of CWN. The results were evaluated with the substitution tests. Substitution test is commonly used in linguistic literature (Tsai et al., 2002); EuroWordnet provided linguistic tests for each semantic relation to examine the validity. In (Tsai et al., 2002), sentence formulae were created following the frame in EuroWordnet to examine the validity of certain semantic relations in Chinese. Linguistic semantic tests help researcher check if two word meanings have a certain kind of semantic relation or not, and further ensure the quality and consistence of the database. Therefore, following the previous framework, a set of sentence formulae based on properties of troponymy was created to verify the correctness of hypernymy-troponymy verb pairs. However, due to data sparseness, the system can achieve only high precision but low recall.

3 Morpho-semantic Linkage

Instead of assuming any *external context* in which words to be linked appear, we propose to exploit the *language-internal evidence* manifested at the morpho-syntactic levels in Chinese, which is assumably guided by underlying semantic composition of morphemes.

3.1 Morpho-semantics in WordNets

The idea of exploiting morpho-semantic information for the enrichment of WordNet has been discussed and implemented in the WordNet community for a while. (Miller and Fellbaum, 2003) first described the importance of adding "morphosemantic links" to WordNet, with later works (Fellbaum et al., 2009) on the classification of regular polysemous patterns of morphosemantic V-N pairs related via *-er* affixation (e.g., *build-builder*).

The notion of *morpho-semantic links* (MSLs) has been applied to other

(morphologically-rich) languages such as Czech (Pala and Hlaváčková, 2007) (in terms of `D-relations`), Turkish (Bilgin et al., 2004) and Bantu languages (Bosch et al., 2008). It is worth of mentioning that the proposed *morpho-semantic relations* or *derivational relations* are relations that hold among literals (lemmas) rather than synsets, which leaves some room of discussion about the extra level these relations should be anchored because neither *paradigmatic* nor *syntagmatic* relations would fit.

It is note here that for morphologically-poor languages like Chinese, the MSLs are quite different in that they do not exist between *stems* and *suffixes*, but between *word-to-be*/*word-used-to-be* morphemes instead. This has the practical advantages for the enrichment of existing paradigmatic relations, as we will introduce in the following.

3.2 Probing Morpho-Semantic Relations in Chinese

The vast majority of Chinese characters represent the *morphemes*. It has been always a controversy over the notion of *wordhood* in the lexical history of Chinese. In a way any Chinese character can be seen as *word-to-be* or *word-used-to-be* morphemes. Given the fact that the relative predominance of the monosyllabic *word* in ancient Chinese has shifted to bi-syllabic words in modern Chinese, the huge semantic weight carried by the morphemes has made the idea of *character-centered* lexicon deeply ingrained in Chinese mind. Orthographically, the lack of word delimiter (such as space) in texts worsens the achievement of consensus regarding the distinction between words, compounds and phrases, and thus makes the segmentation a long-standing heated topic in Chinese NLP.

We follow the cognitive-functional stance in the respect that lexicon and syntax form a continuum rather than two strictly separated modules. We argue that the *Morpho-Semantic Relations* (MSRs), i.e., the ways morphemes combine to form composite meanings, can function as the organic linkage in revealing the composition mechanism among the continuum of different lexical units in varied contexts. In terms of WordNet's paradigmatic relations, this means that morpho-semantic information in Chinese can be used to identify these relations based on the *position* and *semantic role* of morphemes in modification.

In the case of Verb-Verb (compound) words, where the word is composed of two verbal morphemes, linguistics have sorted out different types resulting from the interplay of morphemes within (Li and Thompson, 1981). For instance, for the type of so-called 'parallel' VV compounds, V_1 (verb in the first position) and V_2 (verb in the second position) share the similar meaning (`near synonyms`), such as *bang-zhu* 'help-assist' (help), *fang-qi* 'loosen-abandon' (give up). With a fine-grained sense analysis, we can label the `troponymy` between V_1 and V_1V_2, where V_1 is widely recognized as the component that carries heavier semantic load in VV compound (a.k.a. *left-headedness*).

In the case of Noun-Noun (compound) words, e.g., *noodle-shop* 'mian-dian' (noodle shop), where the word is composed of two nominal morphemes, the N_1*modifier* - N_2*head* structure is prevalently observed (a.k.a. *right-headedness*). The linkage between N_1N_2 and $N2$ can be labeled as `hypernymy-hyponymy`.

4 From MSL to Lexical Semantic Relations

4.1 Hypothesis

As argued in previous section, *Morpho-Semantic Linkage* abound in abundant relational knowledge. In this study, we aim to enrich the CWN with relations leveraged by operationalizing MSL.

The automatic labeling of the lexical semantic relations on word-pairs is quite straightforward. For N_1N_2 compounds, $\prec N_1N_2, N_2\succ$ pairs are labeled with `hypernymy-hyponymy`, and $\prec N_1N_2, N_1\succ$ pairs are labeled with `meronymy-holonymy`.

The cases of VV compounds are trickier, the flow of judgement is shown in algorithm 1. When V_1 has `synonymy` or `near-synonymy` with V_2, then V_1V_2 are troponyms of both V_1 and V_2. If V_2 is on the list of 完住掉開壞成, which is a subclass of the VV compounds that are often called *resultative compounds*, for there is a *causal relation* between the event represented by the first compound of such a compound and the event/state represented by the second component.

Data: VV compounds
Result: Labeled relations between V_1V_2 and V_1/V_2
initialization (POS tagging);
if *V1 is V2* **then**
  return troponymy;
else
  if *V2 is* 完住掉開壞成 **then**
    return causality;
  if *V2 is* 上下來去進回出落入向往過起 **then**
    return directional;
  end
  else
    return pertainymy;
  end
end

Algorithm 1: Pseudo code for relations labeling between V_1V_2 and $V_1/V2$

4.2 Experiments

In this section, we discuss the experiment we designed, the evaluation and error analysis.

The first step is to create a list of term pairs, which a total of 561,703 words covered in CWN [2], Sinica BOW [3], and Ministry of Education Online Chinese Dictionary [4]. In this experiment, we focus only on bi-syllabic words represented by two characters, which constitute the largest proportion of Chinese vocabulary repository.

In order to filter out a coarse-grained bi-syllabic word list, only both characters of a bi-syllabic word that could be found in the big word list, are preserved. Additionally, four principles are applied to construct a more fine-grained word list: [1] the part-of-speech tags of both characters within a bi-syllabic word should be NN or VV; [2] bi-syllabic words containing metaphors are excluded; [3] bi-syllabic morphemic word (e.g., 齷齪 (sordid)) or archaic words (e.g.,搗家) are not included; and [4] proper nouns, (e.g., 成龍(Jackie Chan)) are not considered. Therefore, a list with 1482 bi-syllabic words are produced. Using the hypotheses proposed in section 4.1, the relations are automatically labelled on the related word pairs.

A manual evaluation of the resulting semantic relations lists was conducted. We have created a wiki-based collaborative platform[5] on which registered users can contribute to CWN by adding new entries, editing existing ones and rating one another's contribution to ensure the quality of collective intelligence (Lee et al, 2013). Figure 1 shows the snapshot of the system.

With three linguistic graduate students judging the correctness, the inter-annotator agreement measured by `Fleiss kappa` (Fleiss, 1971) was used, which is defined as:

$$k = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$$

where the numerator expresses the degree of agreement actually achieved, and the denominator the degree of agreement that is possible above chance. As a result, it's interesting to see that there is a very poor agreement between three raters ($k = -0.7069972$) on the predicted relations of $\prec W_1 - W_1W_2 \succ$, which also gets low precision rate; while agreement achieves a moderate degree (with $k = 0.5835113$) on the predicted relations of $\prec W_2 - W_1W_2 \succ$, which also gets high performance in precision.[6] Figure 2 shows the enrichment of relations through the experiment.

4.3 Discussion

The experiment we carried out gives rise to some issues for discussion. Table 1 shows the performance for each predicted relation. When we scrutinize the portion with low precision rate, we found that the problematic cases are mostly from the predicted `meronymy-holonymy` relations between NN compounds, i.e., $\prec N_1N_2 - N_1 \succ$. It is in fact not surprising in that the definition of *part-whole* is not easily stated, and the judgement criteria in the previous literature are not unproblematic too. For instance, given the restrictive rules

[2]See `http://lope.linguistics.ntu.edu.tw/cwn/`
[3]See `http://bow.sinica.edu.tw/`
[4]See `http://dict.revised.moe.edu.tw/`
[5]See `http://lope.linguistics.ntu.edu.tw/cwikin/`
[6]The results will be accessible at `http://140.112.147.131/`

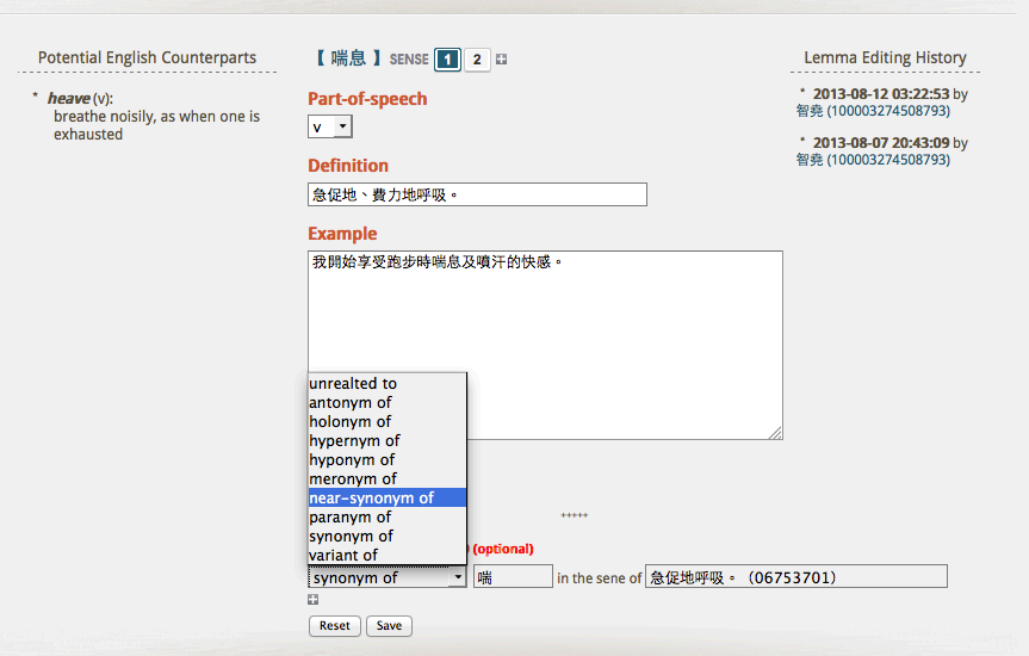

Figure 1: User graphical interface of CWIKIN

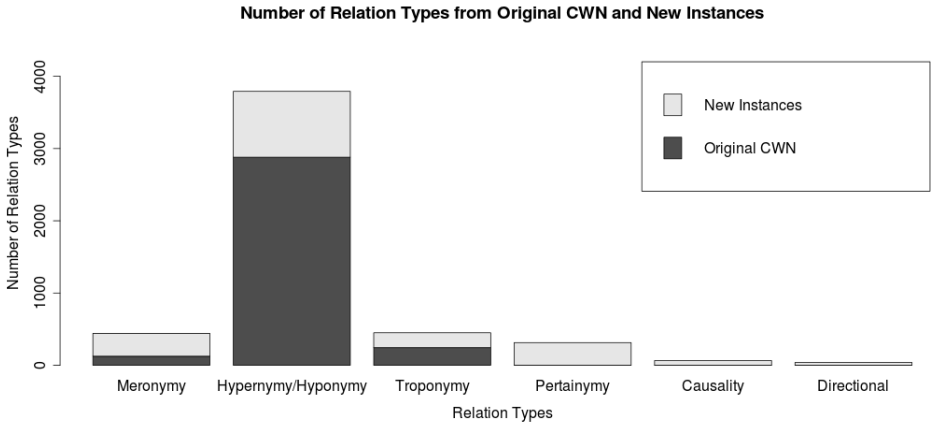

Figure 2: Relations added

that Cruse (1986) sets on the meronymy relation with the co-existence of both the 'N_1N_2 is part of N_1' and 'N_1 has N_1N_2' paraphrases, the raters did not all agree that the relation hold between 黨部 (party headquarter) and 黨 (political party).

Another main error sources come from the predicted `troponymy-hypernymy` relations between $\prec V_1V_2 - V_1 \succ$. Recall that we hypothesize that if V_1 and V_2 are synonymous, then V_1V_2 is automatically labeled as troponym of V_1. The errors arose here can be mainly ascribed to the lack of consistent Chinese thesaurus. In this experiment, the CWN synset (fine-grained synonym determination) and CILIN semantic class (coarse-grained synonym determination) are integrated for prediction, both has different criteria regarding the sameness or nearness of senses between two verbs. In addition, no proper rules for the evaluation of troponymy among raters constitute the difficulties as well.

Furthermore, there are two points can be made. [1], the experiment of relation discovery is conducted at the level of word-lemma, not concept(word-sense), in terms of wordnet, the generic label 'semantic relations' are regarded as the relation occurring between *linguistic units* rather than between concepts (i.e., *synsets*.) Currently, the predicted relations are presumably connected with the first sense of the word lemma in CWN. A fine-grained annotation will be left for future work. [2], in the evaluation task, when the raters did not agree with the predicted relation type, they also provide proper relation types for the pair, which are not *named relations* explicitly defined in WordNet. For example, the *qualia modification* between certain N_1N_2 and N_2, such as 肉醬(meat sauce) - 醬(sauce). This is different from patterned-based approaches where a *bottom-up* methodology is taken because named and explicitly defined semantic relations of interest are presumed before lexico-syntactic patterns are extracted and utilized to search for instances of the relations

5 Conclusion

Lexical semantic relations offers rich linguistic and conceptual knowledge information and are the most to fill in for wordnets. Semantic relations extraction has been one of the most important tasks in many fields. The challenges pertaining to this task are multifaceted. The most active *pattern-based* approaches provide a reasonable solution, but poses difficulties as well.

In this paper, we have presented a *linguistic alternative* to the task in Chinese by resorting to resources of language in itself. Rather than focusing on the *patterns design - relation extraction* model, a notion of *Morpo-semantic links* is proposed to support the extraction and labeling of a wide variety of semantic relations in Chinese. The experiment shows that it is possible to discover semantic relations without being influenced by corpus size and genres. This simple strategy can also serve as the linguistic baseline for related works.

Future works include: [1] extending to VN and NV compounds (Song and Qiu, 1981), and more fined-grained classification of semantic relations among these word-pairs, and [2] mapping with Japanese Wordnet where an amount of Chinese characters are employed for advanced cross-linguistic validation. We also hope that the work presented here will shed new light on the understanding of morpho-semantic representation of natural languages.

References

Alain Auger and Caroline Barrière (eds). 2010. *Probing Semantic Relations*. John Benjamins Publishing, Amsterdam/Philadelphia.

Orhan Bilgin, Özlem Çetinoglu and Kemal Oflazer. 2004. Morphosemantic Relations In and Across Wordnets: A Study Based on Turkish. In: *Proceedings of the Second Global WordNet Conference*, 60–66.

Sonja Bosch, Christiane Fellbaum and Karel Pala. 2008. Enhancing WordNets with Morphological Relations: A Case Study from Czech, English and Zulu. In: *Proceedings of the Fourth Global WordNet Conference*, 74–90.

Philipp Cimiano, Aleksander Pivk Lars, Schmidt-Thieme and Steffen Staab. 2005. Learning Taxonomic Relations from Heterogeneous Sources of

Word Pairs	Type of Relations	**Precision**	**Observations**
W1-W1W2	Meronymy	33%	956
	Pertainymy	25%	352
	Troponymy	29%	174
W2-W1W2	Causality	90%	73
	Directional	90%	43
	Hypernymy/Hyponymy	95%	956
	Pertainymy	93%	241
	Troponymy	92%	169

Table 1: Inter-annotator agreement across all relations

Evidence. In: Buitelaar (eds),*Ontology Learning from Text: Methods, Evaluation and Applications*, 55–73.

Christiane Fellbaum, Anne Osherson, and Peter. E. Clark. 2009. Putting Semantics into WordNet's "Morphosemantic" Links. In: Zygmunt Vetulani and Hans Uszkoreit (eds). *Human Language Technology. Challenges of the Information Society*, 350-358.

Joseph. L. Fleiss. 1971. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76(5): 378–382.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In: *Proceedings of the Fourth International Conference on Computational Linguistics (COLING)*, 539–545.

Shu-Kai Hsieh. 2009. Formal Description of Lexical Semantic Relations. *Concentric: Studies in Linguistics*, 35(1):87–109.

Chu-Ren Huang, I-Ju Tseng, and Dylan Tsai. 2002. Translating Lexical Semantic Relations. In: *SEMANET '02 Proceedings of the 2002 workshop on Building and using Semantic Networks*, 1–7.

Chu-Ren Huang,Shu-Kai Hsieh,Jia-Fei Hong,Yun-Zhu Chen,I-Li Su,Yong-Xiang Chen and Sheng-Wei Huang. 2010. Chinese Wordnet: Design, Implementation, and Application of an Infrastructure for Cross-Lingual Knowledge Processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Chih-Yao Lee, Yu-Yun Chang, Shu-Kai Hsieh, Jia-Fei Hong and Chu-Ren Huang. 2013. *CWIKIN: A wiki that Helps Quicken the Development of Chinese Wordnet.* The 8th International Conference of the Asian Association for Lexicography. Bali, Indonesia.

Charles N. Li and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar.* Berkeley: University of California Press.

Chiao-Shan Lo, Yi-Rung Chen, Chih-Yu Lin, and Shu-Kai Hsieh. 2008. Automatic Labeling of Troponymy for Chinese Verbs. In: *Proceedings of the 20th Conference on Computational Linguistics and Speech Processing (ROCLING).*

George Miller and Christiane Fellbaum. 2003. Morphosemantic Links in WordNet. *Traitement Automatique de Langue*, 44(2):69–80.

Karel Pala and Dana Hlaváčková. 2007. Derivational Relations in Czech WordNet. In: *ACL '07 Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies.* Stroudsburg, PA, USA.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In: *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06).*

Gerardo Sierra, Rodrigo Alarcón, César Aguilar and Carme Bach. 2010. Definitional Verbal Patterns for Semantic Relation Extraction. In: Alain Auger and Caroline Barrière (eds). *Probing Semantic Relations.* John Benjamins Publishing, Amsterdam/Philadelphia.

Zuoyan Song and Qiu Likun. 2013. Qualia Relations in Chinese Nominal Compounds Containing Verbal Elements. *International Journal of Knowledge and Language Processing*, 28(1):114–133.

Dylan B. Tsai, Chu-Ren Huang, Shu-Chuan Tseng, Elanna J.I.Lin, Keh-jiann Chen and Yuan-hsun Chuang. 2002. Chinese Lexical Semantic Relations: Definition and Classification Criteria. *Journal of Chinese Information Processing*, 2002(4):21–31.